Introduction and Motivation
Introduction and motivation for a graduate electrical engineering course on Signal Theory.

Introduction and Motivation

Most areas of electrical and computer engineering (beyond signal processing) deal with signals. Communications is about transmitting, receiving, and interpreting signals. Signals are used to probe and model systems in control and circuit design. The images acquired by radar systems and biomedical devices are signals that change in space and time, respectively. Signals are used in microelectronic devices to convey digital information or send instructions to processors.

This course will provide a mathematical framework to handle signals and operations on signals. Some of the questions that will be answered in this course include:

- What is a signal? How do we represent it?
- How do we represent operations on signals?
- What does it mean for signals to be similar/different from each other?
- When is a candidate signal a good/bad approximation (i.e., a simplified version) of a target signal?
- When is a signal “interesting” or “boring”?
- How can we characterize groups of signals?
- How do we find the best approximation of a target signal in a group of candidates?

Course Overview

Signal Theory

The signal theory presented in this course has three main components:

- *Signal representations and signal spaces*, which provide a framework to talk about sets of signal and to define signal approximations.

- *Distances and norms* to evaluate and compare signals. Norms provide a measure of strength, amplitude, or “interestingness” of a signal, and distances provide a measure of similarity between signals.
- *Projection theory and signal estimation* to work with signals that have been distorted, aiming to recover the best approximation in a defined set.

Operator Theory

Operators are mathematical representations of systems that manipulate a signal. The operator theory presented in this course has three main components:

- *Operator properties* that allow us to characterize their effect on signals in a simple fashion.
- *Operator characterization* that allow us to model their effect on arbitrary inputs.
- *Operator operations* (no pun intended) that allow us to create new systems and reverse the effect of a system on a signal.

Optimization Theory

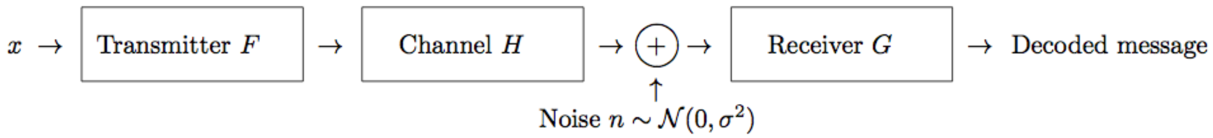
Optimization is an area of applied mathematics that, in the context of our course, will allow us to determine the best signal output for a given problem using defined metrics, such as signal denoising or compression, codebook design, and radar pulse shaping. The optimization theory presented in this course has three main components:

- *Optimization guarantees* that rely on properties of the metrics and signal sets we search over to formally ensure that the optimal signal can be found.
- *Unconstrained optimization*, where we search for the optimum over an entire signal space.

- *Constrained optimization*, where the optimal signal must meet additional specific requirements.

Example

As an example, consider the following communications channel:


$x \rightarrow$
Transmitter F
$\rightarrow$
Channel H
$\rightarrow \oplus \rightarrow$
Receiver G
$\rightarrow$ Decoded message
Noise $n \sim \mathcal{N}(0, \sigma^2)$

Block diagram for a communications channel

A mathematical formulation of this channel requires us to:

- establish which signals x can be input into the transmitter;
- how the transmitter F, the channel H, and the receiver G are characterized;
- how the concatenation of the blocks F and H is expressed;
- how the noise addition operation is formulated;
- how we measure whether the decoded message $\widehat{x}$ is a good approximation of the input x;
- how is the receiver G designed to be optimal for all the choices above.

For this example, by the end of the course, you will be able to solve the problem of selecting the transmitter/receiver pair F, G that minimizes the power of the error $e = \widehat{x} - x$ while meeting maximum transmission power constraints $\frac{\text{power}(F(x))}{\text{power}(x)} < P_{\text{max}}$.

Mathematical Review
Review of background mathematical concepts for the Signal Theory course.

Basic Set Theory

We begin with a quick review of basic set theory from undergraduate courses.

Definition 1 A *set* is an unordered collection of objects denoted by a capital letter A and written explicitly by listing its elements $A = \{a_1, a_2, ...\}$.

Definition 2 The *union* of two sets A and B is denoted by $A \cup B := \{x : x \in A \ \vee x \in B\}$. The *intersection* of two sets A and B is denoted by $A \cap B := \{x : x \in A \wedge x \in B\}$.

Definition 3 A set A is *contained* in another set B, denoted $A \subset B$, if $x \in A \Rightarrow x \in B$. Two sets A and B are *equal* if $x \in A \Leftrightarrow x \in B$.

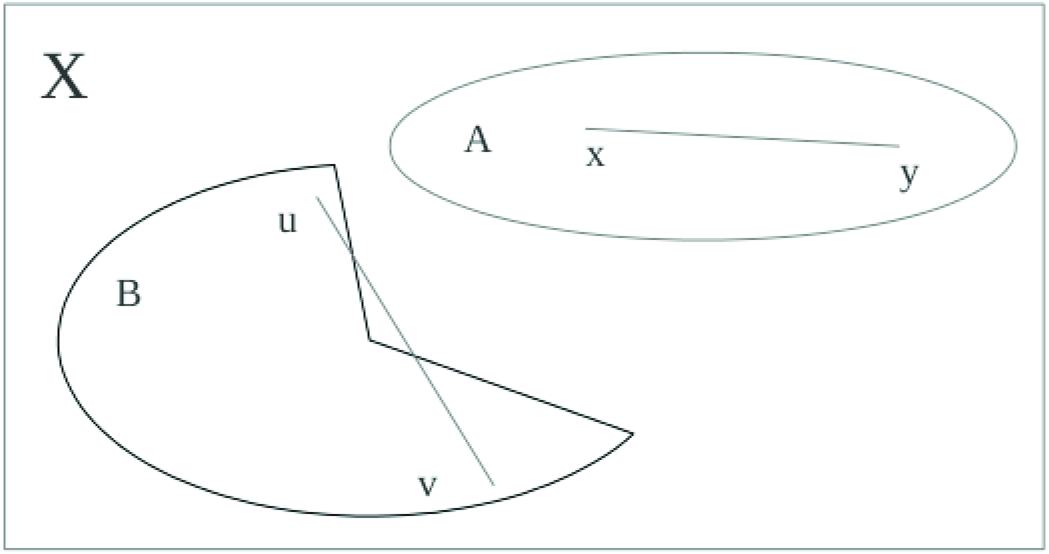
Definition 4 The *complement* of A is the set $\widetilde{A} := \{x : x \notin A\}$. The *empty set* is denoted by $\varphi := \{\}$.

Definition 5 A set is *finite* if it has a finite number of elements. A set is *countably infinite* if there is a one-to-one relationship between its elements and the integers $\mathbb{Z}$. A set is *uncountably infinite* if it is not finite or countably infinite.

Convexity

Definition 6 A set $A \subseteq X$ is *convex* if for all $x, y \in A$ all convex combinations of x and y are in A, i.e., for all $0 \leq \alpha \leq 1$ we have $\alpha x + (1 - \alpha)y \in A$.

Example 1 [link] below shows that the line xy (containing all convex combinations of x and y) is included in A; since this is true for each $x, y \in A$ then the set A is convex. Conversely, for the set B we can find two points $u, v \in B$ such that the line uv is not completely contained in B; therefore, B is not convex.

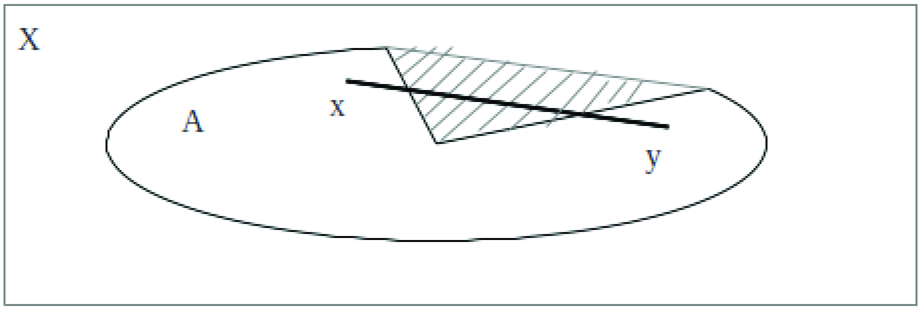


Examples of convex and nonconvex regions.

Fact 1 If A is convex then $\beta A = \{\beta X : x \in A\}$ is convex for $\beta \geq 0$.

Definition 7 The *convex hull* of a set $A \subseteq X$ is the smallest convex set S such that $A \subseteq S$

Example 2 Consider the set A in [link] below, which is not convex. By adding to A all points that are convex combinations of elements of A but not in A (i.e., the points in the shaded region), we obtain the convex hull of A.



Examples of convex and nonconvex regions.

A set with a single element is always convex.

Metric Spaces
Description of signal spaces and metric spaces.

Signal Spaces

We start the content of our course by defining its main concepts of a signal and a signal space.

Definition 1 A *signal* is the value of some quantity as a function of time, space, frequency, etc.; each signal is labeled by a lower-case letter x.

Definition 2 A *signal space* is a set of signals defined by some criterion, labeled by an upper-case letter X (since it is a set).

Some familiar sets of signals are $X=\mathbb{R}$, $X=\mathbb{C}$, and the set of vectors $X=\mathbb{R}^n$.

Definition 3 The signal space $L_2\left[a,b\right]$ contains all signals $x(t)$ such that $x(t)=0$ for all $t<a$ or $t>b$ and $\int_a^b \left|x\left(t\right)\right|^2 dt<\infty$ (i.e., at no time the signal is infinite).

Metric Spaces

Definition 4 A *metric* $d: X\times X\to\mathbb{R}$ is a function used to measure distance between pairs of elements of X with the following properties: for all $x,y,z\in X$,

1. $d(x,y)=d(y,x)$ (symmetry),
2. $d(x,y)\geq 0$ (non-negativity),
3. $d(x,y)=0 \Leftrightarrow x=y$,
4. $d(x,z)\leq d(x,y)+d(y,z)$ (triangle inequality).

If d is a metric on X, the pair (X,d) is called a *metric space*. A set X can have multiple metrics, leading to different metric spaces.

Example 1 The following are some initial examples of metric spaces:

- $X = \mathbb{R}$ with $d_0\left(x, y\right) = \left|x - y\right|$ for all $x, y \in \mathbb{R}$: it is easy to check properties (1-4).
- $X = \mathbb{R}$ with $d'\left(x, y\right) = \begin{cases} 1 & \text{if} \;\; x \neq y \\ 0 & \text{if} \;\; x = y \end{cases}$ for all $x, y \in \mathbb{R}$: it is easy to check properties (1-3). To verify (4), assume that $d(x, y) + d(y, z) = 0$; then both $d(x, y) = 0$ and $d(y, z) = 0$, which means $x = y$ and $y = z$; by transitivity, $x = z$ and $d(x, z) = 0 \leq d(x, y) + d(y, z)$, as desired. Now assume that $d(x, y) + d(y, z) = 1$; then we immediately get $d(x, z) \leq d(x, y) + d(y, z)$, as desired.
- $X = \mathbb{R}^n$ with metric $d_2^n\left(x, y\right) := \left(\sum_{i=1}^n \left|x_i - y_i\right|^2\right)^{1/2}$, known as the Euclidean metric.
- $X = \mathbb{R}^n$ with metric $d_1^n\left(x, y\right) := \left(\sum_{i=1}^n \left|x_i - y_i\right|\right)$.
- $X = L_2\left[a, b\right]$ with metric $d_2\left(x, y\right) := \left(\int_a^b \left|x\left(t\right) - y\left(t\right)\right|^2 dt\right)^{1/2}$. Formally, d_2 is a *pseudometric* on X, since there are signals $x \neq y$ that yield $d_2\left(x, y\right) = 0$. However, one can define a new signal space where all signals with $d(x, y) = 0$ are equal to each other.
- $X = L_2\left[a, b\right]$ with metric $d_p\left(x, y\right) := \left(\int_a^b \left|x\left(t\right) - y\left(t\right)\right|^p dt\right)^{1/p}$, for $1 \leq p < \infty$, an extension of the metric d_2.
- $X = L_2\left[a, b\right]$ with metric $d_\infty\left(x, y\right) := \mathrm{sup}_{t \in [a,b]} \left|x\left(t\right) - y\left(t\right)\right|$; this metric solves the equivalence problem of d_2.

Here, sup $\left(A\right)$ is the *supremum* of A, i.e., the smallest value $x_0 \in \mathbb{R}$ such that $x \leq x_0 \;\forall\; x \in A$. Similarly, inf $\left(A\right)$ is the *infimum* of A, i.e., the largest value x_0 such that $x_0 \leq x \;\forall\; x \in A$.

Convergent Sequences, Cauchy Sequences, and Complete Spaces
Description of convergent sequences and Cauchy sequences in metric spaces. Description of complete spaces.

Convergence

The concept of convergence evaluates whether a sequence of elements is getting “closer” to a given point or not.

Definition 1 Assume a metric space (X,d) and a countably infinite sequence of elements $\{x_n\} := \{x_n, n = 1,2,3,...\} \subseteq X$. The sequence $\{x_n\}$ is said to *converge* to $x \in X$ if for any $\epsilon > 0$ there exists an integer $n_0 \in \mathbb{Z}^+$ such that $d(x,x_n) < \epsilon$ for all $n \geq n_0$. A convergent sequence can be denoted as $\lim_{n\to\infty} x_n = x$ or $x_n \overset{n\to\infty}{\longrightarrow} x$.

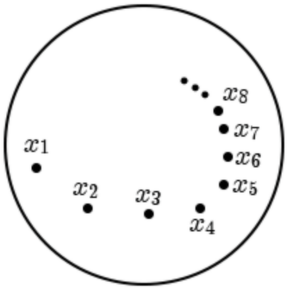


Illustration of a convergent sequence $\{x_i\}$.

Note that in the definition n_0 is implicitly dependent on ϵ, and therefore is sometimes written as $n_0\left(\epsilon\right)$. Note also that the convergence of a sequence depends on both the space X and the metric d: a sequence that is convergent in one space may not be convergent in another, and a sequence that is convergent under some metric may not be convergent under another. Finally, one can abbreviate the notation of convergence to $x_n \to x$ when the index variable n is obvious.

Example 1 In the metric space $(\mathbb{R}, d_0)$ where $d_0\left(x,y\right) = |x-y|$, the sequence $x_n = 1/n$ gives $x_n \overset{n\to\infty}{\longrightarrow} 0$: fix ϵ and let $n_0 > \lceil 1/\epsilon \rceil$ (i.e., the smallest integer that is larger than $1/\epsilon$). If $n \geq n_0$ then
Equation:

$$d_0\left(0,x_n\right) = |0{-}x_n| {=} |x_n| {=} x_n = \frac{1}{n} \leq \frac{1}{n_0} < \frac{1}{\lceil 1/\epsilon \rceil} \leq \frac{1}{1/\epsilon} = \epsilon,$$

verifying the definition. So by setting $n_0\left(\epsilon\right) > \lceil 1/\epsilon \rceil$, we have shown that $\{x_n\}$ is a convergent sequence.

Example 2 Here are some examples of non-convergent sequences in $(\mathbb{R}, d_0)$:

- $x_n = n^2$ diverges as $n \to \infty$, as it constantly increases.
- $x_n = 1 + \left(-1\right)^n$ (i.e., the sequence $\{x_n\} = \{0,2,0,2,...\}$) diverges since for $\epsilon < 1$ there does not exist an n_0 that holds the definition for any choice of limit x. More explicitly, assume that a limit x exists. If $x \notin [0,2]$ then for any $\epsilon \leq 2$ one sees that for either even or odd values of n we have $d(x,x_n) > \epsilon$, and

so no n_0 holds the definition. If $x \in [0,2]$ then select $\epsilon = \frac{1}{2}\ \min\ (x, 2-x)$. We will have that $d(x, x_n) \geq \epsilon$ for all n, and so no n_0 can hold the definition. Thus, the sequence does not converge.

Theorem 1 If a sequence converges, then its limit is unique.

Proof: Assume for the sake of contradiction that $x_n \to x$ and $x_n \to y$, with $x \neq y$. Pick an arbitrary $\epsilon > 0$, and so for the two limits we must be able to find n_0 and n_0', respectively, such that $d(x, x_n) < \epsilon/2$ if $n > n_0$ and $d(y, x_n) < \epsilon/2$ if $n > n_0'$. Pick n^* $>$max (n_0, n_0'); using the triangle inequality, we get that $d\left(x,y\right) \leq d\left(x, x_{n^*}\right) + d\left(x_{n^*}, y\right) < \epsilon/2 + \epsilon/2 = \epsilon$. Since for each ϵ we can find such an n^*, it follows that $d(x,y) < \epsilon$ for all $\epsilon > 0$. Thus, we must have $d(x,y) = 0$ and $x = y$, and so the two limits are the same and the limit must be unique.

Cauchy Sequences

The concept of a Cauchy sequence is more subtle than a convergent sequence: each pair of consecutive elements must have a distance smaller than or equal than that of any previous pair.

Definition 2 A sequence $\{x_n\}$ is a *Cauchy sequence* if for any $\epsilon > 0$ there exists an $n_0 \in \mathbb{Z}^+$ such that for all $j, k \geq n_0$ we have $d(x_j, x_k) < \epsilon$.

As before, the choice of n_0 depends on ϵ, and whether a sequence is Cauchy depends on the metric space (X, d). That being said, there is a connection between Cauchy sequences and convergent sequences.

Theorem 2 Every convergent sequence is a Cauchy sequence.

Proof: Assume that a sequence $x_n \to x$ in (X, d). Fix $\epsilon > 0$. Since $\{x_n\}$ is convergent, there must exist an $n_0 \in \mathbb{Z}^+$ such that $d(x_n, x) < \epsilon/2$ for all $n \geq n_0$. Now, pick $j, k \geq n_0$. Then, using the triangle inequality, we have $d\left(x_j, x_k\right) \leq d\left(x_j, x\right) + d\left(x, x_k\right) < \epsilon/2 + \epsilon/2$ (since both j and k are greater or equal to n_0) and so $d(x_j, x_k) < \epsilon$. Therefore, the sequence $\{x_n\}$ is Cauchy.

One may wonder if the opposite is true: is every Cauchy sequence a convergent sequence?

Example 3 Focus on the metric space (X, d_0) with $X = (0,1] = \{x \in \mathbb{R} : 0 < x \leq 1\}$. Now consider the sequence $x_n = 1/n$ in X. This is a Cauchy sequence: one can show this by picking $n_0\left(\epsilon\right) > \lceil 2/\epsilon \rceil$ and using the triangle inequality to get
Equation:

$$d_0\left(x_j, x_k\right) \leq d_0\left(x_j, 0\right) + d_0\left(x_k, 0\right) = |x_j| + |x_k| = \frac{1}{j} + \frac{1}{k} \leq \frac{1}{n_0} + \frac{1}{n_0} = \frac{2}{n_0} < \frac{2}{\lceil 2/\epsilon \rceil} \leq \frac{2}{2/\epsilon} = \epsilon.$$

However, this is not a convergent sequence: we've shown earlier that $x_n \to 0$. Since $0 \notin (0,1]$ and a sequence has a unique limit, then there is no $x \in (0,1]$ such that $x_n \to x$.

Complete Metric Spaces

Whether Cauchy sequences converge or not underlies the concept of completeness of a space.

Definition 4 A *complete metric space* is a metric space in which all Cauchy sequences are convergent sequences.

Example 5 The metric space (X, d_0) with $X = (0,1]$ from earlier is not complete, since we found a Cauchy sequence that converges to a point outside of X. We can make it complete by adding the convergence point, i.e., if $X' = [0, 1]$ then (X', d_0) is a complete metric space.

Example 6 Let $C[T]$ denote the space of all continuous functions with support T. If we pick the metric $d_2(x, y) = \left(\int_T |x(t) - y(t)|^2 dt \right)^{1/2}$, then (X, d_2) is a metric space; however, it is not a complete metric space.

To show that the space is not complete, all we have to do is find a Cauchy sequence of signals within $C[T]$ that does not converge to any signal in $C[T]$. For simplicity, fix $T = [-1, 1]$. Fortuitously, we find the sequence illustrated in [link], which can be written as

Equation:

$$x_n(t) = \begin{cases} -1 & \text{if } t < -1/n, \\ 1 & \text{if } t > 1/n, \\ nt & \text{if } -1/n \leq t \leq 1/n. \end{cases}$$

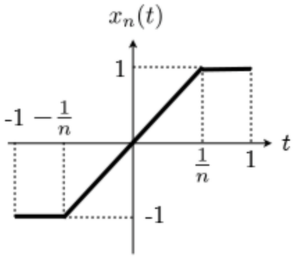


Illustration of a convergent sequence $\{x_i\}$.

Since all these functions are continuous and defined over T, then $\{x_n\}$ is a sequence in $C[T]$. We can show that $\{x_n\}$ is a Cauchy sequence: let n_0 be an integer and pick $j \geq k \geq n_0$. Then,

Equation:

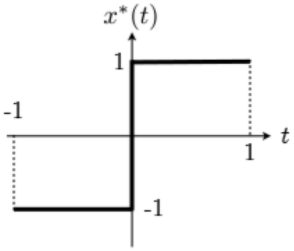
$$\begin{aligned} d_2(x_j, x_k) &= \left(\int_{-1}^{1} |x_k(t) - x_j(t)|^2 dt \right)^{1/2} = \left(\int_{-1/j}^{1/j} |x_k(t) - x_j(t)|^2 dt \right)^{1/2} \leq \left(\int_{-1/j}^{1/j} 1^2 dt \right)^{1/2} \\ &\leq (2/j)^{1/2} \leq \sqrt{2/n_0}. \end{aligned}$$

So for a given $\epsilon > 0$, by picking n_0 such that $\sqrt{2/n_0} < \epsilon$ (say, for example, $n_0 > \lceil 2/\epsilon^2 \rceil$), we will have that $d_2(x_j, x_k) < \epsilon$ for all $j, k > n_0$; thus, the sequence is Cauchy. Now, we must show that the sequence does not converge within $C[T]$: we will find a point $x^* \notin C[T]$ such that for $X' = C[T] \cup \{x^*\}$ the sequence $x_n \to x^*$ in (X', d_2). By inspecting the sequence of signals, we venture the guess

Equation:

$$x^*(t) = \begin{cases} -1 & \text{if } \ t < 0, \\ 1 & \text{if } \ t > 0,, \\ 0 & \text{if } \ t = 0. \end{cases}$$

illustrated in [link].



Initial guess for a convergence point x^* for the sequence $\{x_i\}$.

For this signal, we will have

Equation:

$$\begin{aligned} d_2\left(x_n, x^*\right) &= \left(\int_{-1}^{1} \left|x_n(t) - x^*(t)\right|^2 dt\right)^{1/2} = \left(\int_{-1}^{0} \left|x_n(t) - x^*(t)|^2 dt + \int_{0}^{1} \left|x_n(t) - x^*(t)\right|^2 dt\right)^{1/2} \\ &= \left(\int_{-1/n}^{0} |nt - (-1)|^2 dt + \int_{0}^{1/n} |nt - 1|^2 dt\right)^{1/2} \\ &= \left(\int_{-1/n}^{0} |nt + 1|^2 dt + \int_{0}^{1/n} |1 - nt|^2 dt\right)^{1/2} \\ &= \left(\int_{0}^{1/n} |1 - nt|^2 dt + \int_{0}^{1/n} |1 - nt|^2 dt\right)^{1/2} = \left(2\int_{0}^{1/n} (1 - nt)^2 dt\right)^{1/2} \\ &= \left(\frac{2}{3n}\right)^{1/2}. \end{aligned}$$

So if we select n_0 such that $\left(\frac{2}{3n_0}\right)^{1/2} < \epsilon$, i.e., $n_0 > \frac{2}{3\epsilon^2}$, then we have that $d_2\left(x_n, x^*\right) < \epsilon$ for $n > n_0$, and so we have shown that $x_n \to x^*$. Now, since a convergent sequence has a unique limit and $x^* \notin C[T]$, then $\{x_n\}$ does not converge in $(C[T], d_2)$ and this is not a complete metric space.

The property of equivalence between Cauchy sequences and convergent sequences often compels us to define metric spaces that are complete by choosing the metric appropriate to the signal space. For example, by switching the distance metric to $d_\infty(x, y) = \sup_{t\in T} |x(t) - y(t)|$, the metric space $(C[T], d_\infty)$ becomes complete.

Continuity and Convergence of Functions
Definitions of continuity and convergence for functions defined on metric spaces.

Continuity for Functions

Definition 1 A function $f : (X, d_x) \to (Y, d_y)$ is continuous at a point $x_0 \in X$ if: for any $\epsilon > 0$ there exists a $\delta > 0$ such that if $d_x\left(x_0, x_1\right) < \delta$, then $d_y\left(f\left(x_0\right), f\left(x_1\right)\right) < \epsilon$.

Definition 2 A function: $f : (X, d_x) \to (Y, d_y)$ is uniformly continuous if: for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $x_0 \in X$: if $d_x\left(x_0, x_1\right) < \delta$, then $d_y\left(f\left(x_0\right), f\left(x_1\right)\right) < \epsilon$.

The difference between these definitions is that for a function to be simply continuous at a point, one need only find a δ for the given input x_0, while for a function to be uniformly continuous, one needs to find for a given ϵ a single value of δ that works in the definition for every point over which the function is defined.

Example 1 Consider the function $f : (C[T], d_\infty) \to (\mathbb{R}, d_0)$ defined as $f(x) = x\left(t_0\right)$. In words, the input to f is a continuous function over T, and the output from f is the value of the input function evaluated at $t = t_0$. We ask the question: *Is f continuous at some function input $x_1(t)$?*

- Designate a pair of functions $x_1(t), x_2(t)$ such that: $d_\infty\left(x_1(t), x_2(t)\right) < \delta$, where
  Equation:

$$d_\infty\left(x_1(t), x_2(t)\right) = \sup_{t \in T} \left|x_1(t) - x_2(t)\right|.$$

  In other words, $sup_{t \in T} \left|x_1(t) - x_2(t)\right| < \delta$.
- Next, we see that $d_0\left(f\left(x_1\right), f\left(x_2\right)\right) = \left|x_1\left(t_0\right) - x_2\left(t_0\right)\right|$.
- Because $\left|x_1\left(t_0\right) - x_2\left(t_0\right)\right| \leq \sup_{t \in T} \left|x_1(t) - x_2(t)\right|$, by the definition of the supremum, we can simply select $\delta = \epsilon$ to get that if $d_0\left(f\left(x_1\right), f\left(x_2\right)\right) < \delta$, then

Equation:

$$d_0\left(f\left(x_1\right), f\left(x_2\right)\right) \leq d_\infty\left(x_1\left(t\right), x_2\left(t\right)\right) < \delta = \epsilon.$$

This shows the continuity of $f(x)$ at $x_1\left(t\right)$. However, because the selection of δ did not depend on the value of x_1, we have also shown that the function f is uniformly continuous.

Convergence of Functions

Definition 3 The sequence of functions $\{f_n\}$, $f_n : (X, d_x) \rightarrow (Y, d_y)$ converges pointwise to $f : (X, d_x) \rightarrow (Y, d_y)$ if: for each $x \in X$, the sequence of values $\{f_n\left(x\right)\}$ converges to $f(x)$ in (Y, d_y).

Definition 4 The sequence of functions $\{f_n\}$, $f_n : (X, d_x) \rightarrow (Y, d_y)$ converges uniformly to $f : (X, d_x) \rightarrow (Y, d_y)$ if: for each $\epsilon > 0$, there exists $n_0 \in \mathbb{Z}^+$ such that if $n \geq n_0$, then $d_y\left(f\left(x\right), f_n\left(x\right)\right) < \epsilon$ for all $x \in X$.

The difference between these definitions is that for uniform convergence, there must exist a single value of n_0 that works in the definition of continuity for all possible values of $x \in X$.

Example 2 Consider the sequence of functions $x_n\left(t\right) : ([0, 1], d_0) \rightarrow ([0, 1], d_0)$ given by $x_n\left(t\right) = \frac{t}{n}$. One naturally suspects that $x_n\left(t\right)$ may be converging to the zero-valued function. We can show this formally as follows:

- Pick some t_0, and check if $\{x_1\left(t_0\right), x_2\left(t_0\right), x_3\left(t_0\right), x_4\left(t_0\right), \ldots\}$ is converging to 0: Denote $a_n = x_n\left(t_0\right) = \frac{t_0}{n}$, and notice that $d_0\left(a_n, 0\right) = |a_n - 0| = \frac{t_0}{n}$. So, to pick an n_0 such that $a_n < \epsilon$ if $n \geq n_0$, one only needs to note that since $\frac{t_0}{n} \leq \frac{t_0}{n_0}$, it then suffices for $\frac{t_0}{n_0} < \epsilon$, or in other words $n_0 > \frac{t_0}{\epsilon}$. So this sequence converges pointwise to 0.

- Additionally, because the range of possible inputs to $x(t)$ is $t \in [0, 1]$, we could select $n_0 > \frac{1}{\epsilon}$. Because 1 is the maximum value for t_0, $n_0 > \frac{1}{\epsilon} > \frac{t_0}{\epsilon}$ will work for all values of $t_0 \in [0, 1]$, and so the sequence $\{x_n\}$ also converges uniformly to the zero function.

Vector Spaces
Description of vector spaces, subspaces, bases and spans.

Vector Spaces

Definition 1 A *linear vector space* $(X, R, +, \cdot)$ is given by a signal space X (called vectors), a set of scalars R, an addition operation $+ : X \times X \to X$, and a multiplication operation $\cdot : R \times X \to X$, such that:

1. X forms a group under addition:

   a. $\forall\ x, y \in X \quad \exists!\ x + y \in X,$ (closed under addition)
   b. $\exists\ 0 \in X$ such that $0 + X = X + 0 = X.$ (additive identity)
   c. $\forall\ x \in X \quad \exists\ y \in X$ such that $x + y = 0,$ (additive inverse)
   d. $\forall\ x, y, z \in X \quad x + (y + z) = (x + y) + z.$ (associative law)

2. Multiplication has the following properties: for any $x, y \in X$ and $a, b \in R$:

   a. $a \cdot x \in X,$ (closure in X under multiplication)
   b. $a \cdot (b \cdot x) = (a \cdot b) \cdot x,$ (compatibility)
   c. $(a + b) \cdot x = a \cdot x + b \cdot x,$ (distributive law over R)
   d. $a \cdot (x + y) = a \cdot x + a \cdot y.$ (distributive law over X)

3. The set R has the following properties:

   a. There exists $1 \in R$ s.t. $1 \cdot x = x \quad \forall\ x \in X,$ (multiplication identity)
   b. There exists $0 \in R$ s.t. $0 \cdot x = 0 \quad \forall\ x \in X.$ (multiplicative null element)

Example 1 Here are some examples of vector spaces:

- $X = \mathbb{R}^n$ (space of all vectors of length n) over $R = \mathbb{R}$ is a vector space.
- $X = \mathbb{C}^n$ ($\mathbb{C}$ is complex numbers) over $R = \mathbb{C}$ is a vector space.
- $X = \mathbb{R}^n$ over $R = \mathbb{C}$ is not a vector space, because closure in X under multiplication is not met.
- $X = C[T]$ (continuous functions in T) over $R = \mathbb{R}$ is a vector space.

Subspaces

Definition 2 A subset $M \subseteq X$ is a *linear subspace* of X if M itself is a linear vector space. Note that, in particular, this implies that any subspace M must obey $0 \in M$.

Example 2 Here are some examples of subspaces:

- In $X = \mathbb{R}^2$ over $R = \mathbb{R}$, any line that passes through the origin is a subspace of X:

Equation:

$$M = \Big\{(x, y) \in \mathbb{R}^2 \ \ \text{such that} \ \ \frac{y}{x} = c\Big\}.$$

- In $X = C[T]$ over $R = \mathbb{R}$, the followings are subspaces of X:
  Equation:

$$\begin{aligned} M_1 &= \big\{f(x) = ax^2 + bx + c : a, b, c \in \mathbb{R}\big\} \\ M_2 &= \{f(x) : f(x_0) = 0\}. \end{aligned}$$

  In contrast, the set $M_3 = \{f(x) : f(x_0) = a \neq 0\}$ is not a subspace.

Proposition 1 If M and N are subspaces of X, then $M \cap N$ is also a subspace.

Proof: We assume that M and N hold properties of linear vector space, and show that so does $M \cap N$:

1. **Equation:**

$$x, y \in M \cap N \quad \Rightarrow \quad \begin{Bmatrix} x, y \in M & \Rightarrow & x + y \in M \\ x, y \in N & \Rightarrow & x + y \in N \end{Bmatrix} \quad \Rightarrow \quad x + y \in M \cap N$$

2. **Equation:**

$$\begin{Bmatrix} M & \text{linear vector space} \Rightarrow 0 \in M \\ N & \text{linear vector space} \Rightarrow 0 \in N \end{Bmatrix} \Rightarrow 0 \in M \cap N$$

3. **Equation:**

$$x \in M \cap N \Rightarrow \begin{Bmatrix} x \in M & \exists y \in M & \text{s.t.} & x + y = 0 \\ x \in N & \exists y \in N & \text{s.t.} & x + y = 0 \end{Bmatrix} \Rightarrow y \in M \cap N$$

The other properties are shown in a similar fashion.

Definition 3 A vector $x \in X$, where $(X, R, +, \cdot)$ is a vector space, is a *linear combination* of a set $\{x_1, x_2, ..., x_n\} \subseteq X$ if it can be written as $x = \sum_{i=1}^{n} a_i \cdot x_i$, $a_i \in R$. The set of all linear combinations of a set of points $\{x_1, x_2, ..., x_n\}$ builds a linear subspace of X.

Example 3 $Q = \sum_{i=0}^{2} a_i x^i$ is a linear subspace of $(C[T], \mathbb{R}, +, \cdot)$ containing the set of all quadratic functions, as it corresponds to all linear combinations of the set of functions $\big\{x^2, x, 1\big\}$.

Bases and Spans

Definition 4 For the set $S = \{x_1, x_2, ..., x_n\} \subseteq X$, the *span* of S is written as
Equation:

$$[S] = \mathrm{span}\,(S) = \left\{x : x = \sum_{i=1}^{n} a_i x_i, a_i \in R\right\}.$$

Example 4 The space of quadratic functions Q is written as $Q = [S_1]$, with $S_1 = \left\{x^2, x, 1\right\}$. The space can also be written as $[S_2]$ with $S_2 = \left\{1, x, x^2 - 2\right\})$, i.e. $[S_1] = [S_2]$. To prove this, we need to show $[S_2] \subseteq [S_1]$ and $[S_1] \subseteq [S_2]$. For the former case we have
Equation:

$$\begin{aligned} x &= a_1 + a_2 x + a_3 \left(x^2 - 2\right) \\ &= (a_1 - 2a_3) + a_2 x + a_3 x^2, \end{aligned}$$

which means that every element that can be spanned by S_2, can also be spanned by S_1, and hence $[S_2] \subseteq [S_1]$. The latter case can be shown in a similar manner.

Definition 5 A set S is a *linearly independent set* if
Equation:

$$\sum_{i=1}^{n} a_i x_i = 0 \Leftrightarrow a_i = 0, \forall\ i \in \{1, 2, ..., n\}.$$

Otherwise, the set S is *linearly dependent*.

Definition 6 A finite set S of linearly independent vectors is a *basis* for the space X if $[S] = X$, i.e. if X is spanned by S.

Definition 7 The *dimension* of X is the number of elements of its basis $|S|$. A vector space for which a finite basis does not exist is called an *infinite-dimensional space*.

Theorem 1 Any two bases of a subspace have the same number of elements.

Proof: We prove by contradiction: assume that $S_1 = \{x_1, ..., x_n\}$ and $S_2 = \{y_1, ..., y_m\}$, $m > n$, are two bases for a subspace X with different numbers of elements. We have that since $y_1 \in X$ it can be written as a linear combination of S_1:
Equation:

$$y_1 = \sum_{i=1}^{n} a_i x_i.$$

Order the elements of S_1 above so that a_1 is nonzero; since y_1 must be nonzero then at least one such a_i must exist. Solving the above equation for x_1 yields

Equation:

$$x_1 = \frac{1}{a_1}\left(y_1 - \sum_{i=2}^{n} a_i x_i\right).$$

Thus $\{y_1, x_2, x_3, ..., x_n\}$ is a basis, in terms of which we can write any vector of the space X, including y_2:

Equation:

$$y_2 = b_1 y_1 + \sum_{i=2}^{n} b_i x_i.$$

Since y_1, y_2 are linearly independent, at least one of the values of $b_i,\ i > 2$, must be nonzero. Sort the remaining x_i so that b_2 is nonzero. Solving for x_2 results in

Equation:

$$x_2 = \frac{1}{b_2} y_2 - \frac{b_1}{b_2} y_1 + \sum_{i=3}^{n} \frac{b_i}{b_2} x_i.$$

Therefore, $\{y_1, y_2, x_3, ..., x_n\}$ is a basis for X. Continuing in this way, we can eliminate each x_i, showing that $\{y_1, y_2, ..., y_n\}$ is a basis for X. Thus, we have $y_{n+1} = \sum_{i=1}^{n} c_i y_i$, or equivalently:

Equation:

$$c_{n+1} y_{n+1} + \sum_{i=1}^{n} c_i y_i = 0 \quad \text{with} \quad c_{n+1} = -1.$$

As a result, S_2 is linearly dependent and is not a basis. Therefore, all bases of X must have the same number of elements.

Basis Representations

Having a basis in hand for a given subspace allows us to express the points in the subspace in more than one way. For each point $x \in [S]$ in the span of a basis $S = \{S_1, S_2 ... S_n\}$, that is,

Equation:

$$x = \sum_{i=1}^{n} a_i S_i,$$

there is a one-to-one map (i.e., an equivalence) between $x \in [S]$ and $a = \{a_1, ..., a_n\} \in R^n$, that is, both x and a uniquely identify the point in S. This is stated more formally as a theorem.

Theorem 2 If S is a linearly independent set, then

Equation:

$$\sum_{i=1}^{n} a_i S_i = \sum_{i=1}^{n} b_i S_i$$

if and only if $a_i = b_i$ for $i = 1, 2 ... n$.

Proof: Theorem 1 states that the scalars $\{a_1, ..., a_n\}$ are unique for x. We begin by assuming that indeed

Equation:

$$\sum_{i=1}^{n} a_i S_i = \sum_{i=1}^{n} b_i S_i.$$

This implies

Equation:

$$\begin{aligned} \sum_{i=1}^{n} a_i S_i - \sum_{i=1}^{n} b_i S_i &= 0, \\ \sum_{i=1}^{n} (a_i - b_i) S_i &= 0. \end{aligned}$$

Since the elements of S are linearly independent, each one of the scalars of the sum must be zero, that is, $a_i - b_i = 0$ and so $a_i = b_i$ for each $i = 1, ..., n$.

Example 5 (Digital Communications) A transmitter sends two waveforms:
Equation:

$$S_1\left(t\right)=\sqrt{2/T}\cos\left(2\pi f_c t\right)\quad t\in\left[0,T\right]\quad\text{if bit 1 is transmitted,}$$

Equation:

$$S_0\left(t\right)=\sqrt{2/T}\sin\left(2\pi f_c t\right)\quad t\in\left[0,T\right]\quad\text{if bit 0 is transmitted.}$$

The signal $r(t)$ recorded by the receiver is continuous, that is, $r(t)\in C[T]$. Assuming that the propagation delay is known and corrected at the receiver, we will have they the received signal must be in the span of the possible transmitted signals, i.e., $r\left(t\right)\in\text{span}\left(S_1\left(t\right),S_0\left(t\right)\right)$. One can check that $S_1\left(t\right)$ and $S_2\left(t\right)$ are linearly independent. Thus, one can use a unique choice of coefficients a_0 and a_1 that denote whether bit 0 or bit 1 is transmitted and contain the amount of attenuation caused by the transmission:
Equation:

$$r\left(t\right)=a_1 S_1\left(t\right)+a_0 S_0(t).$$

The uniqueness of this representation can only be obtained if the transmitted signals $S_0\left(t\right)$ and $S_1\left(t\right)$ are linearly independent. The waveforms above are used in in phase shift keying (PSK); other similar examples include frequency shift keying (FSK) and quadrature amplitude modulation (QAM).

Normed Spaces
Description of norms, normed spaces, and Banach spaces

Distances and metrics allow us to evaluate how different two signals are from each other. Norms allow us to evaluate how “big”, “important”, or “interesting” a given signal is.

Definition 1 Assume X is a linear vector space. A *norm* on X is a function $\| \cdot \|$: $X \rightarrow \mathbb{R}$ with the following properties for all $x, y \in X$ and $a \in R$:

1. $\| x \| \geq 0$, (non-negativity)
2. $\| x \| = 0$ if and only if $x = 0$ (zero norm for zero vector)
3. $\| a \cdot x \| = |a| \cdot \| x \|$, (scaling)
4. $\| x + y \| \leq \| x \| + \| y \|$. (triangle inequality)

$\| x \|$ is read as the norm of x or length of x.

Intuitively, one can say that $\| x \|$ is the distance between x and the zero vector (more on this soon).

Definition 2 A vector space X with a norm $\| \cdot \|$ is called a *normed linear vector space*$(X, \| \cdot \|)$ (or a normed space for brevity).

Definition 3 Let $(X, \| \cdot \|)$ be a normed space. The *induced metric* or *induced distance* is given by $d_I\left(x, y\right) = \| x - y \|$.

Definition 4 If a normed space is complete under the induced metric, then it is called a *Banach space*.

All norms induce distances, but not all distances are induced by norms.

Example 1 Consider the distance
Equation:

$$d'\left(x, y\right) = \begin{cases} 0 & \text{if } \; x = y, \\ 1 & \text{if } \; x \neq y. \end{cases}$$

Let us assume that there exists a norm $\parallel x \parallel_i = d' (x, 0)$ that would induce this distance. We would then have for $x \neq 0$ and $\alpha \notin \{-1, 0, 1\}$ that $\parallel \alpha x \parallel_i = d' (\alpha x, 0) = 1$ and $\parallel x \parallel_i = d' (x, 0) = 1$, which contradicts $\parallel \alpha x \parallel_i = |\alpha| \parallel x \parallel_i$. Thus $\parallel \cdot \parallel_i$ is not a valid norm.

In contrast, here are some examples of valid norms.

Example 2 The vector space $X = C[T]$ accepts the norm $\parallel x \parallel_\infty = \text{sup}_{t\in T} |x (t)|$. The induced distance is $d_i (x, y) = \text{sup}_{t\in T} |x (t) - y (t)| = d_\infty (x, y)$; it is straightforward to prove properties (1–4). We previously showed that the metric space $(C [T], d_\infty)$ is complete, and so $(C [T], \parallel \cdot \parallel_\infty)$ is a Banach space.

Example 3 The vector space $X = \mathbb{R}^n$ accepts the norm $\parallel x \parallel_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$. The induced metric is $d_i (x, y) = \parallel x - y \parallel_2 = \left(\sum_{i=1}^n |x_i - y_i|^2\right)^{1/2} = d_2 (x, y)$, the Euclidean distance. Thus $\parallel \cdot \parallel_2$ is known as the Euclidean norm. The spaces $(\mathbb{R}^n, d_2)$ are Banach spaces for all values of n.

Example 4 The vector space $X = [0, 1)$ accepts the norm $\parallel x \parallel = |x|$. The induced metric $d_i (x, y) = |x - y| = d_0 (x, y)$ is the standard metric for the reals. Since we have previously shown that (X, d_0) is not a complete vector space, then the space $((0, 1], \parallel \cdot \parallel)$ is not a Banach space.

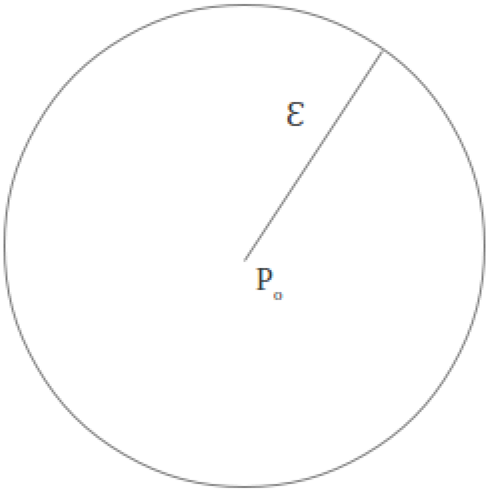
Topological Concepts
Review of basic topological sets

Interiors and Closures

Definition 1 Let (X,d) be a metric space. The *ball* $B(x_0,\epsilon)$ centered at $x_0 \in X$ and of radius $\epsilon \geq 0$ is defined as
$B\left(x_0,\epsilon\right) = \{x \in X : d\left(x,x_0\right) < \epsilon\}$.

If $(X,\|\cdot\|)$ is a metric space, then $B\left(x_0,\epsilon\right) = \{x \in X :\| x{-}x_0 \|< \epsilon\}$.

Example 1 In the metric space $\left(\mathbb{R}^2,d_2\right)$, the ball $B(p_0,\epsilon)$ is a circle centered at p_0 and of radius ϵ, as illustrated in [link].



A euclidean/ℓ_2 ball with center p_0 and radius ϵ

The following four definitions are fundamental in the study of topology.

Definition 2 Let $P \subset X$ where (X,d) is a metric space. A vector $p_0 \in P$ is an *interior point* of P if there exists $\epsilon > 0$ such that $B(p_0,\epsilon) \subseteq P$.

Intuitively, the notion of an interior point is a point that is not in the “boundary” of the set, as a ball around it is contained within the set.

Definition 3 The *interior* of a set P is the collection of all the interior points of P and is denoted as P°.

Intuitively, closure points are points that are arbitrarily close to the set P; note however that a closure point need not be in P, but only have a sequence of elements of P that converge to it.

Definition 4 A point $p_1 \in X$ is a *closure point* of P if for all $\epsilon > 0$ we have that $B(p_1, \epsilon) \cap P \neq \emptyset$.

Definition 5 The *closure* of a set P is the set of all closure points of P denoted as $\overline{P}$.

Open and Closed Sets

Topology is the study of open and closed sets, defined below.

Definition 6 A set P is said to be *open* if $P = P^\circ$, i.e., every point in P is an interior point of P.

Definition 7 A set P is said to be *closed* if $P = \overline{P}$, i.e., P contains all its closure points.

Fact 1 Since all interior points of P are in P and every point in P is a closure point of P, we have that $P^\circ \subseteq P \subseteq \overline{P}$.

Example 2 The following are examples of open and closed sets.

- The set $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ is closed.
- The set $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ is open. To show this, we must show that every point $x \in (a, b)$ is an interior point. Pick an arbitrary $x \in (a, b)$, and define $\epsilon =$min $\left(\frac{x-a}{2}, \frac{b-x}{2}\right)$. Then the ball $B\left(x, \epsilon\right) = \{u \in \mathbb{R} : |u - x| < \epsilon\}$ can be rewritten as the set of all

points u such that $-\epsilon + x < u < \epsilon + x$. Using the definition of ϵ, we have that if $u \in B(x, \epsilon)$ then
Equation:

$$-\frac{x-a}{2} + x \leq u \leq \frac{b-x}{2} + x,$$

or equivalently,
Equation:

$$\frac{x+a}{2} \leq u \leq \frac{b+x}{2}.$$

Since $a < x < b$, we have
Equation:

$$a < \frac{x+a}{2} \leq u \leq \frac{b+x}{2} < b,$$

and so $u \in (a, b)$. Since $u \in B(x, \epsilon)$ was arbitrary, we then have $B(x, \epsilon) \subseteq (a, b)$ and x is an interior point of (a, b). Now since $x \in (a, b)$ was arbitrary, then the set (a, b) is open.

Properties of Open and Closed Sets

Theorem 1 (i) If A is open then A^C is closed. (ii) If A is closed then A^C is open.

Proof: (i) We will prove by contradiction: Assume A is open and A^C is not closed, that is, there exists a closure point x of A^C such that $x \notin A^C$, that is, $x \in A$. Since x is a closure point of A^C, we have that for every $\epsilon > 0$,
Equation:

$$B\left(x, \epsilon\right) \cap A^C \neq \emptyset.$$

Since A is open and $x \in A$, then x is an interior point of A, which means that there exists $\epsilon_0 > 0$ such that $B(x, \epsilon_0) \subseteq A$, which means that

$B\left(x,\epsilon_0\right)\cap A^C=\emptyset$, a contradiction with [link]. Therefore, we must have that A^C is closed.

(ii) Assume A is closed, which means that A contains all its interior points. That means that if $x\in A^C$ then x is not a closure point of A, meaning that there for some $\epsilon_0>0$ we must have $B(x,\epsilon_0)\cap A=\emptyset$. This means that $B\left(x,\epsilon_0\right)\subseteq A^C$, and so x is an interior point of A^C. Since x was an arbitrary point in A^C, this means that A^C is open.

Proposition 1 The intersection of a finite number of open sets is open, and the union of an arbitrary collection of open sets is open.

Proof: We will limit the proof to two sets, which can be extended in each case using a proof by induction argument.

We first show that if A_1,A_2 are open then $A_1\cap A_2$ is open, i.e., $\left(A_1\cap A_2\right)^O=A_1\cap A_2$. Assume $x\in A_1\cap A_2$; then $x\in A_1$ and $x\in A_2$. Since A_1,A_2 are open then there exists ϵ_1,$\epsilon_2>0$ such that $B\left(x,\epsilon_1\right)\subseteq A_1$ and $B\left(x,\epsilon_2\right)\subseteq A_2$. Set $\epsilon=min(\epsilon_1,\epsilon_2)$; then, $B\left(x,\epsilon\right)\subseteq B\left(x,\epsilon_1\right)$ and $B\left(x,\epsilon\right)\subseteq B\left(x,\epsilon_2\right)$. By transitivity of inclusion, we have that $B\left(x,\epsilon\right)\subseteq A_1$ and $B\left(x,\epsilon\right)\subseteq A_2$. Theferore, $B\left(x,\epsilon\right)\subseteq A_1\cap A_2$.

Next, we show that if A_1,A_2 are open then $A_1\cup A_2$ is open, i.e., $\left(A_1\cup A_2\right)^O=A_1\cup A_2$. Assume $x\in A_1\cup A_2$; then $x\in A_1$ or A_2. If $x\in A$, then there exists $\epsilon_1>0$ s.t. $B\left(x,\epsilon_1\right)\subseteq A_1\subseteq A_1\cup A_2$. Similarly, if $x\in A_2$, there exists $\epsilon_2>0$ s.t. $B\left(x,\epsilon_2\right)\subseteq A_2\subseteq A_1\cup A_2$. So $x\in A_1\cup A_2$ is an interior point of $A_1\cup A_2$ and therefore $\left(A_1\cup A_2\right)^O=A_1\cup A_2$.

Proposition 2 The union of a finite number of closed sets is closed. The intersection of an arbitrary collection of closed sets is closed.

The following useful properties are proven in “Optimization by Vector Space Methods” by David Luenberger, pages 25 and 38.

Proposition 3 If C is convex then its interior C^O and closure $\overline{C}$ are convex.

Proposition 4 A subset of a Banach space is complete if and only if it is closed.

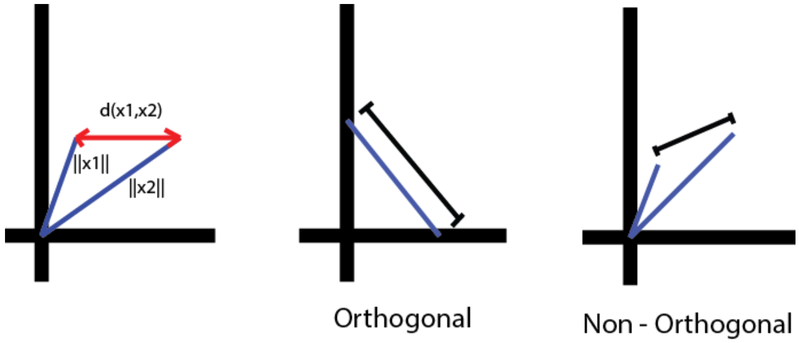
Proposition 5 Any finite dimensional subspace of a normed linear space is complete.

Why are Banach spaces useful? In optimization, we want to show that if an increasingly better solution can be found then an optimum must exist.

Inner Product Spaces and Hilbert Spaces
Review of inner products and inner product spaces.

Inner Products

We have defined distances and norms to measure whether two signals are different from each other and to measure the “size” of a signal. However, it is possible for two pairs of signals with the same norms and distance to exhibit different behavior - an example of this contrast is to pick a pair of orthogonal signals and a pair of non-orthogonal signals, as shown in [link].



An example of orthogonality in a two-dimensional space: distances and norms are not indicative of orthogonality; two pairs of vectors with the same distance can have arbitrary angle between them.

To obtain a new metric that distinguishes between orthogonal and non-orthogonal we use the *inner product*, which provides us with a new metric of “similarity”.

Definition 1 An *inner product* for a vector space $(X, R, +, \cdot)$ is a function $\langle\cdot,\cdot\rangle : X \times X \to R$, sometimes denoted $(\cdot|\cdot)$, with the following properties: for all $x, y, z \in X$ and $a \in R$,

1. $\langle x, y\rangle$ = $\langle y, x\rangle$ (complex conjugate property),
2. $\langle x + y, z\rangle$ = $\langle x, z\rangle + \langle y, z\rangle$ (distributive property),
3. $\langle \alpha x, y\rangle$ = $\alpha\langle x, y\rangle$ (scaling property),
4. $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0$ if and only if $x = 0$.

A vector space with an inner product is called an *inner product space* or a *pre-Hilbert* space.

It is worth pointing out that properties (2-3) say that the inner product is linear, albeit only on the first input. However, if $R = \mathbb{R}$, then the properties (2-3) hold for both inputs and the inner product is linear on both inputs.

Just as every norm induces a distance, every inner product induces a norm: $||x||_i = \sqrt{\langle x, x\rangle}$.

Hilbert Spaces

Definition 2 An inner product space that is complete under the metric induced by the induced norm is called a *Hilbert space*.

Example 1 The following are examples of inner product spaces:

1. $X = \mathbb{R}^n$ with the inner product $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = y^T x$. The corresponding induced norm is given by $||x||_i = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2} = ||x||_2$, i.e., the ℓ_2 norm. Since $(\mathbb{R}^n, \| \cdot \|_2)$ is complete, then it is a Hilbert space.
2. $X = C[T]$ with inner product $\langle x, y \rangle = \int_T x\,(t) y\,(t) dt$. The corresponding induced norm is $||x||_i = \sqrt{\int_T x(t)^2 dt} = ||x||_2$, i.e., the L_2 norm.
3. If we allow for $X = C[T]$ to be complex-valued, then the inner product is defined by $\langle x, y \rangle = \int_T x\,(t) y(t) dt$, and the corresponding induced norm is $||x||_i = \sqrt{\int_T x\,(t)\;x(t) dt} = \sqrt{\int_T |x\,(t)|^2 dt} = ||x||_2$.
4. $X = \mathbb{C}^n$ with inner product $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = y^H x$; here, x^H denotes the Hermitian of x. The corresponding induced norm is $||x||_i = \sqrt{\sum_{i=1}^{n} |x_i|^2} = ||x||_2$.

Theorem 1 (Cauchy-Schwarz Inequality) Assume X is an inner product space. For each $x, y \in X$, we have that $|\langle x, y \rangle| \leq ||x||_i ||y||_i$, with equality if (i) $y = ax$ for some $a \in R$; (ii) $x = 0$; or (iii) $y = 0$.

Proof: We consider two separate cases.

- if $y = 0$ then $\langle x, y \rangle = \langle y, x \rangle = \langle 0 \cdot y, x \rangle = 0 \langle y, x \rangle = 0\, \langle x, y \rangle = 0 = \parallel x \parallel_i \parallel y \parallel_i$. The proof is similar if $x = 0$.
- If $x, y \neq 0$ then $0 \leq \langle x - ay, x - ay \rangle = \langle x, x \rangle - a\, \langle y, x \rangle - a\, \langle x, y \rangle + aa\, \langle y, y \rangle$, with equality if $x - ay = 0$, i.e., $x = ay$ for some $a \in R$. Now set $a = \frac{\langle x, y \rangle}{\langle y, y \rangle}$, and so $a = \frac{\langle y, x \rangle}{\langle y, y \rangle}$. We then have
  Equation:

  $$\begin{aligned} 0 \quad & \leq \langle x, x \rangle - \frac{\langle x, y \rangle}{\langle y, y \rangle} \langle y, x \rangle - \frac{\langle y, x \rangle}{\langle y, y \rangle} \langle x, y \rangle + \frac{\langle x, y \rangle}{\langle y, y \rangle} \frac{\langle y, x \rangle}{\langle y, y \rangle} \langle y, y \rangle \\ & \leq \langle x, x \rangle - \frac{\langle x, y \rangle\, \langle x, y \rangle}{||y||^2} = ||x||^2 - \frac{|\langle x, y \rangle|^2}{||y||^2}. \end{aligned}$$

  This implies $\frac{|\langle x, y \rangle|^2}{||y||^2} \leq ||x||^2$, and so since all quantities involved are positive we have $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$.

Properties of Inner Products Spaces

In the previous lecture we discussed norms induced by inner products but failed to prove that they are valid norms. Most properties are easy to check; below, we check the triangle inequality for the induced norm.

Lemma 1 If $\parallel x \parallel_i = \sqrt{\langle x,x\rangle}$, then $\parallel x+y \parallel_i \leq \parallel x \parallel_i + \parallel y \parallel_i$.

From the definition of the induced norm,
Equation:

$$\begin{aligned} \|x+y\|_i^2 &= \langle x+y, x+y\rangle, \\ &= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle, \\ &= \|x\|_i^2 + \langle x,y\rangle + \langle x,y\rangle + \|y\|_i^2 \\ &= \|x\|_i^2 + 2\,\mathrm{real}\left(\langle x,y\rangle\right) + \|y\|_i^2. \end{aligned}$$

At this point, we can upper bound the real part of the inner product by its magnitude: $\mathrm{real}\left(\langle x,y\rangle\right) \leq |\langle x,y\rangle|$. Thus, we obtain
Equation:

$$\begin{aligned} \|x+y\|_i^2 &\leq \|x\|_i^2 + 2\left|\langle x,y\rangle\right| + \|y\|_i^2, \\ &\leq \|x\|_i^2 + 2\|x\|_i\|y\|_i + \|y\|_i^2, \\ &\leq \left(\|x\|_i + \|y_i\|\right)^2, \end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality. Thus we have shown that $\|x+y\|_i \leq \|x\|_i + \|y\|_i$. Here's an interesting (and easy to prove) fact about inner products:

Lemma 2 If $\langle x,y\rangle = 0$ for all $x \in X$ then $y = 0$.

Proof: Pick $x = y$, and so $\langle y,y\rangle = 0$. Due to the properties of an inner product, this implies that $y = 0$.

Earlier, we considered whether all distances are induced by norms (and found a counterexample). We can ask the same question here: are all norms induced by inner products? The following theorem helps us check for this property.

Theorem 2 (Parallelogram Law) If a norm $\|\cdot\|$ is induced by an inner product, then $\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$ for all $x, y \in X$.

This theorem allows us to rule out norms that cannot be induced.

Proof: For an induced norm we have $\|x\|^2 = \langle x,x\rangle$. Therefore,
Equation:

$$\begin{aligned}\|x+y\|^2+\|x-y\|^2 &= \langle x+y,x+y\rangle+\langle x-y,x-y\rangle,\\ &= \langle x,x\rangle+\langle x,y\rangle+\langle y,x\rangle+\langle y,y\rangle+\langle x,x\rangle-\langle x,y\rangle-\langle y,x\rangle+\langle y,y\rangle,\\ &= 2\langle x,x\rangle+2\langle y,y\rangle,\\ &= 2\Big(\|x\|^2+\|y\|^2\Big).\end{aligned}$$

Example 2 Consider the normed space $(C\,[T],L_\infty)$, and recall that $\|x\|_\infty = \text{sup}_{t\in T}\,|x\,(t)|$. If this norm is induced, then the Parallelogram law would hold. If not, then we can find a counterexample. In particular, let $T=[0,2\pi]$, $x(t)=1$, and $y(t)=\text{cos}\,(t)$. Then, we want to check if $\|x+y\|^2+\|x-y\|^2=2\left(\|x\|^2+\|y\|^2\right)$. We compute:

Equation:

$$\begin{aligned}\|x\|_\infty &= 1,\\ \|y\|_\infty &= 1,\\ \|x+y\|_\infty &= \|1+\text{ cos }(t)\| = \sup_{t\in T}\,|1+\text{ cos }(t)| = 1+1=2,\\ \|x-y\|_\infty &= \|1-\text{ cos }(t)\| = \sup_{t\in T}\,|1-\text{ cos }(t)| = 1-(-1)=2.\end{aligned}$$

Plugging into the two sides of the Parallelogram law,

Equation:

$$\begin{aligned}2^2+2^2 &= 2\big(1^2+1^2\big),\\ 8 &= 4,\end{aligned}$$

and the Parallelogram law does not hold. Thus, the L_∞ norm is not an induced norm.

The ell_2 space
Formulation of the ell_2 space

Definition 1 We define the ℓ_2 space of infinite sequences of finite energy as
Equation:

$$\ell_2 = \{(x_1, x_2, x_3, ...) \in \mathbb{R}^\infty \text{such that} \sum_{i=1}^{\infty} |x_i|^2 < \infty\}.$$

On this space, define the inner product $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i \overline{y_i}$, and obtain the induced norm $\|x\|_2 = \sqrt{\sum_{i=1}^{\infty} |x_i|^2}$, which we term the ℓ_2 norm.

Here $\mathbb{R}^\infty$ refers to the set of all infinite sequences of real values. Note that this implies that the sequence $|x_i|$ must converge to zero as $i \to \infty$. Note also can then say that the ℓ_2 space consists of all infinite sequences with finite ℓ_2 norm.

Theorem 1 The ℓ_2 space is a Hilbert space.

To prove this theorem, we need the following lemma.

Lemma 1 A Cauchy sequence in a normed space is bounded.

Proof of Lemma 1: Let $\{x_n\}$ be a Cauchy sequence and let n_0 be an integer such that $\| x_n - x_{n_0} \|< 1$ for $n > n_0$. For $n > N$, we have $\| x_n \|=\| x_n - x_{n_0} + x_{n_0} \|\leq\| x_n - x_{n_0} \| + \| x_{n_0} \|\leq 1+ \| x_{n_0} \|$. Now set $M = \max_{1 \leq n \leq n_0} \| x_n \| +1$. Then, $\| x_n \|< M$ for all $n = 1, 2, 3, ...$.

Proof of Theorem 1: We show that ℓ_2 is complete by proving that all Cauchy sequences converge in ℓ_2. Assume that $\{x^{(n)}\}$ is a Cauchy sequence in ℓ_2. Then, if $x^{(n)} = \left(x_1^{(n)}, x_2^{(n)}, ...\right)$, we have that there exists some n_0 such that if $j, k > n_0$, then for each $a = 1, 2, ...$,
Equation:

$$\left|x_a^{(j)} - x_a^{(k)}\right| \leq \sqrt{\sum_{i=1}^{\infty} \left|x_i^{(j)} - x_i^{(k)}\right|^2} = \|x^{(j)} - x^{(k)}\| < \epsilon.$$

Therefore, each sequence $\left\{x_a^{(n)}\right\}$ for each $a = 1, 2, \cdots$ is a Cauchy sequence in the space $(\mathbb{C}, d_0)$, which is complete. Thus, each sequence $\left\{x_a^{(n)}\right\}$ must converge in $\mathbb{C}$ to some value $\overset{*}{x_a}$.

Define $x^* = \left(\overset{*}{x_1}, \overset{*}{x_2}, ...\right)$. We show that $x^* \in \ell_2$. According to Lemma 1, the sequence $\{x_n\}$ is bounded by some constant M. For each pair $a, k = 1, 2, ...,$ we have
Equation:

$$\sum_{i=1}^{k} \left|x_a^{(n)}\right|^2 \leq \|x^{(n)}\|_2^2 \leq M^2.$$

This inequality is valid for each value of n, and so we must have
Equation:

$$\sum_{i=1}^{k} \left|\overset{*}{x_a}\right|^2 \leq M^2.$$

Additionally, his inequality is valid for each value of k, and so we must have
Equation:

$$\sum_{i=1}^{\infty} \left|\overset{*}{x_a}\right|^2 \leq M^2.$$

Thus, we have shown that $x^* \in \ell_2$. The last point we need to show is that $x^{(n)} \to x^*$. First, since the sequence $\left\{x^{(n)}\right\}$ is Cauchy, we have that there exists an n_0 such that if $j, k \geq n_0$ and for each $l = 1, 2, ...$, we have

Equation:

$$\sum_{i=1}^{l} \left| x_i^{(j)} - x_i^{(k)} \right|^2 \leq \|x^{(j)} - x^{(k)}\|_2^2 \leq \frac{\epsilon}{4}.$$

Observe also that for each i there exists $n_{0,i}$ such that if $k \geq n_{0,i}$, then $\left| x_i^{(k)} - \overset{*}{x_i} \right| < \frac{2^{-i}\epsilon}{4}$; therefore, if $k \geq$max $(n_0, \sup_i n_{0,i})$, we have

Equation:

$$\begin{aligned} \sum_{i=1}^{l} \left| x_i^{(j)} - \overset{*}{x_i} \right|^2 & \leq 2 \sum_{i=1}^{l} \left(\left| x_i^{(j)} - x_i^{(k)} \right|^2 + \left| x_i^{(k)} - \overset{*}{x_i} \right|^2 \right) \\ & \leq 2 \left(\sum_{i=1}^{l} \left| x_i^{(j)} - x_i^{(k)} \right|^2 + \sum_{i=1}^{l} \left| x_i^{(k)} - \overset{*}{x_i} \right|^2 \right) \\ & < 2 \left(\frac{\epsilon}{4} + \frac{\epsilon}{4} \sum_{i=1}^{l} 2^{-i} \right) \leq 2 \left(\frac{\epsilon}{4} + \frac{\epsilon}{4} \right) = \epsilon, \end{aligned}$$

where the first inequality come from the fact that $|a + b|^2 \leq 2(|a|^2 + |b|^2)$, which is easy to check. Since this is true for each $l = 1, 2, ...$, then it follows that if $j \geq \overset{*}{n_0}$:=max $(n_0, \sup_i n_{0,i})$, then

Equation:

$$\|x^{(j)} - \overset{*}{x}\| = \sum_{i=1}^{\infty} \left| x_i^{(j)} - \overset{*}{x_i} \right|^2 < \epsilon.$$

This shows that $x^{(n)} \rightarrow x^*$.

Orthogonality
Definition of orthogonal vectors, sets, and subspaces; properties and benefits.

Recall that a set S is a basis for a subspace X if $[S] = X$ and S has linearly independent elements. If S is a basis for X then each $x \in X$ is uniquely determined by $(a_1, a_2, ..., a_n)$ such that $\sum_{i=1}^{n} a_i s_i = x$. In this sense, we could operate either with x itself or with the vector $a = [a_1, ...a_n] = R^n$. One would wonder then whether particular operations can be performed with a representation a instead of the original vector x.

Example 1 Assume $x, y \in X$ have representations $a, b \in R^n$ in a basis for X. Can we say that $\langle x, y \rangle = \langle a, b \rangle$?

For the particular example of $X = L_2\left[0, 1\right]$, $S = \left\{1, t, t^2\right\}$ so that $[S] = Q$, the set of all quadratic functions supported on $[0, 1]$. Pick $x = 2 + t + t^2$ and $y = 1 + 2t + 3t^2$. One can see then that if we label $s_1 = 1$, $s_2 = t$, $s_3 = t^2$, then the coefficient vectors for x and y are $a = [2\ 1\ 1]$ and $b = [1\ 2\ 3]$, respectively. Let us compute both inner products:
Equation:

$$\begin{aligned} \langle x, y \rangle &= \int_0^1 x\left(t\right) y\left(t\right) dt = \int_0^1 \left(2 + t + t^2\right)\left(1 + 2t + 3t^2\right) dt = \frac{187}{20} \approx 9.35, \\ \langle a, b \rangle &= 2 + 2 + 3 = 7. \end{aligned}$$

Since $7 \neq 9.35$, we find that we fail to obtain the desired equivalence between vectors and their representations.

While this example was unsuccessful, simple conditions on the basis S will yield this desired equivalence, plus many more useful properties.

Several definitions of orthogonality will be useful to us during the course.

Definition 1 A pair of vectors x and y in an inner product space are *orthogonal* (denoted $x \perp y$) if the inner product $\langle x, y \rangle = 0$.

Note that 0 is immediately orthogonal to all vectors.

Definition 2 Let X be an inner product space. A set of vectors $S \subseteq X$ is *orthogonal* if $\langle x, y \rangle = 0$ for all $x, y \in S, x \neq y$.

Definition 3 Let X be an inner product space. A set of vectors $S \subseteq X$ is *orthonormal* if S is an orthogonal set and $\| \ s \ \|= \sqrt{\langle s,s\rangle} = 1$ for all $s \in S$.

Definition 4 A vector x in an inner product space X is *orthogonal to a set*$S \subseteq X$ (denoted $x \perp S$) if $x \perp y$ for all $y \in S$.

Definition 5 Let X be an inner product space. Two sets $S_1 \subseteq X$ and $S_2 \subseteq X$ are *orthogonal* (denoted $S_1 \perp S_2$) if $x \perp y$ for all $x \in S_1$ and $y \in s_2$.

Definition 6 The *orthogonal complement* $S^\perp$ of a set S is the set of all vectors that are orthogonal to S.

Benefits of Orthogonality

Why is orthonormality good? For many reasons. One of them is the equivalence of inner products that we desired in a previous example. Another is that having an orthonormal basis allows us to easily find the coefficients $a_1, ...a_n$ of x in a basis S.

Example 2 Let $x \in X$ and S be a basis for X (i.e., $[S]=X$). We wish to find $a_1, ...a_n$ such that $x = \sum_{i=1}^n a_i s_i$. Consider the inner products

Equation:

$$\langle x, s_i\rangle \ = \left\langle \sum_{j=1}^n a_j s_j, s_i \right\rangle = \sum_{j=1}^n a_j \langle s_j, s_i\rangle,$$

due to the linearity of the inner product in the first term. If S is orthonormal, then we have that for $i \neq j$ $\langle s_j, s_i\rangle = 0$. In that case the sum above becomes

Equation:

$$\langle x, s_i\rangle \ = a_i \langle s_i, s_i\rangle = a_i \| \ s_i \ \|^2 = a_i,$$

due to the orthonormality of S. In other words, for an orthonormal basis S one can find the basis coefficients as $a_i = \langle x, s_i\rangle$.

If S is not orthonormal, then we can rewrite the sum above as the product of a row vector and a column vector as follows:

Equation:

$$\langle x, s_i \rangle = \begin{bmatrix} \langle s_1, s_i \rangle & \langle s_2, s_i \rangle & \cdots & \langle s_n, s_i \rangle \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

We can then stack these equations for $i = 1, ..., n$ to obtain the following matrix-vector multiplication:

Equation:

$$\underbrace{\begin{bmatrix} \langle x, s_1 \rangle \\ \langle x, s_2 \rangle \\ \vdots \\ \langle x, s_n \rangle \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} \langle s_1, s_1 \rangle & \langle s_2, s_1 \rangle & \cdots & \langle s_n, s_1 \rangle \\ \langle s_1, s_2 \rangle & \langle s_2, s_3 \rangle & \cdots & \langle s_n, s_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle s_1, s_n \rangle & \langle s_2, s_n \rangle & \cdots & \langle s_n, s_n \rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}}_{a}.$$

The nomenclature given above provides us with the matrix equation $\beta = G \cdot a$, where β and G have entries $\beta_i = \langle x, s_i \rangle$ and $G_{i,j} = \langle s_j, s_i \rangle$, respectively.

Definition 7 The matrix G above is called the *Gram matrix (or Gramian)* of the set S.

In the particular case of orthonormal S, it is easy to see that $G = I$, the identity matrix, and so $a = \beta$ as given earlier. For invertible Gramians G, one could compute the coefficients in vector form as $a = G^{-1}\beta$. For square matrices (like G), invertibility is linked to singularity.

Definition 8 A *singular* matrix is a non-invertible square matrix. A *non-singular matrix* is an invertible square matrix.

Theorem 1 A matrix is singular if $G \cdot x = 0$ for some $x \neq 0$. A matrix is non-singular if $G \cdot x = 0$ only for $x = 0$.

The link between this notion of singularity and invertibility is straightforward: if G is singular, then there is some $a \neq 0$ for which $G \cdot a = 0$. Consider the mapping $y = G \cdot x$; we would also have $y = G(x + a)$. Since $x \neq x + a$, one cannot “invert” the mapping provided by G into y.

Theorem 2 S is linearly independent if and only if G is non-singular (i.e. $Gx = 0$ if and only if $x = 0$).

Proof: We will prove an equivalent statement: S is linearly dependent if and only if G is singular, i.e., if and only if there exists a vector $x \neq 0$ such that $Gx = 0$.

$(\Rightarrow)$ We first prove that if S is linearly dependent then G is singular. In this case there exist a set $\{a_i\} \subseteq R$, with at least one nonzero, such that $\sum_{i=1}^{n} a_i s_i = 0$. We then can write $\left\langle \sum_{i=1}^{n} a_i s_i, s_j \right\rangle = \langle 0, s_j \rangle = 0$ for each s_j. Linearity allows us to take the sum and the scalar outside the inner product:
Equation:

$$\sum_{i=1}^{n} a_i \langle s_i, s_j \rangle = 0.$$

We can rewrite this equation in terms of the entries of the Gram matrix as $\sum_{i=1}^{n} a_i G_{ji} = 0$. This sum, in turn, can be written as the vector inner product
Equation:

$$\begin{bmatrix} G_{j1} & \cdots & G_{1n} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = 0,$$

which is true for every value of j. We can therefore collect these equations into a matrix-vector product:
Equation:

$$\begin{bmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \ddots & \vdots \\ G_{n1} & \cdots & G_{nn} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Therefore we have found a nonzero vector a for which $Ga = 0$, and therefore G is singular. Since all statements here are equalities, we can backtrack to prove the opposite direction of the theorem $(\Leftarrow)$.

Pythagorean Theorem

There are still more nice proper ties for orthogonal sets of vectors. The next one has well-known geometric applications.

Theorem 3 (Pythagorean theorem) If x and y are orthogonal ($x \perp y$), then $\| x \|^2 + \| y \|^2 = \| x + y \|^2$.

Proof:
Equation:

$$\| x + y \|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle$$

Because x and y are orthogonal, $\langle x, y \rangle = \langle y, x \rangle = 0$ and we are left with $\langle x, x \rangle = \| x \|^2$ and $\langle y, y \rangle = \| y \|^2$. Thus: $\| x + y \|^2 = \| x \|^2 + \| y \|^2$.

Gram-Schmidt Process

Description of the Gram-Schmidt procedure to formulate orthonormal bases.

The Gram-Schmidt algorithm or procedure is used to find an orthonormal basis for the subspace $[S]$, even if S is not linearly independent. The algorithm is formally defined as follows:

- **Inputs**: Set of vectors $S = \{s_1, \cdots, s_n\}$.
- **Outputs**: Orthonormal basis elements $W = \{w_1, \cdots, w_n\}$ that span the same space: $[W] = [S]$.
- **Procedure:**

1. Take the first element of the set and divide it by its norm (that is, *normalize* the element):
   Equation:
   $$w_1 = \frac{s_1}{\parallel s_1 \parallel}.$$
2. Take the second element and subtract the projection into the first basis element:
   Equation:
   $$t_2 = s_2 - \langle s_2, w_1 \rangle w_1.$$
   Normalize the result t_2:
   Equation:
   $$w_2 = \frac{t_2}{\parallel t_2 \parallel}.$$
   It is easy to check that w_1 and w_2 are orthonormal:
   Equation:
   $$\begin{aligned} \langle w_1, w_2 \rangle &= \left\langle w_1, \frac{t_2}{\parallel t_2 \parallel} \right\rangle = \frac{1}{\parallel t_2 \parallel} \langle w_1, s_2 - \langle s_2, w_1 \rangle w_1 \rangle \\ &= \frac{1}{\parallel t_2 \parallel} \Big(\langle w_1, s_2 \rangle - \overline{\langle s_2, w_1 \rangle} \langle w_1, w_1 \rangle \Big) = \frac{1}{\parallel t_2 \parallel} \Big(\langle w_1, s_2 \rangle - \overline{\langle s_2, w_1 \rangle} \Big) \\ &= \frac{1}{\parallel t_2 \parallel} \cdot 0 = 0. \end{aligned}$$
   Thus, w_1 and w_2 are orthonormal.
3. The second step is repeated for each additional element; i^{th} element follows the following formula:
   Equation:

$$\begin{aligned} t_i &= s_i - \sum_{j=1}^{i-1} \langle s_i, w_j \rangle w_j, \\ w_i &= \frac{t_i}{\parallel t_i \parallel}. \end{aligned}$$

When the set S includes linearly dependent vectors, some of the unnormalized vectors $t_i = 0$, as the projections will cancel out with some elements of S. As a result, the number of vectors needed will be higher than the dimensionality of the space; in other words, the dimensionality of $[S]$ will be smaller than the cardinality of the set $|S|$.

Example 1 Let $S = \{1, t, t^2\}$, where $s_1 = 1$, $s_2 = t$ and $s_3 = t^2$. We can therefore write the set of quadratic functions as $Q = [S]$, and $S \subseteq C(T)$, where $T = [0, 1]$. Recall that for this space, the inner product is written as $\langle x, y \rangle = \int_0^1 x\,(t) y\,(t) dt$. We obtain a basis for Q using the Gram-Schmidt procedure: it requires us to compute several norms and inner products on the way.

1. Solve for $w_1 :\parallel s_1 \parallel = \sqrt{\langle s_1, s_1 \rangle} = \sqrt{\int_0^1 1 \cdot 1 dt} = 1$, and so $w_1\,(t) = \frac{s_1(t)}{\|s_1\|} = \frac{1}{1} = 1$.
2. Solve for w_2 :
   Equation:

$$\begin{aligned} \langle s_2, w_1 \rangle &= \int_0^1 t \cdot 1 dt = \frac{t^2}{2}\bigg|_{t=[0,1]} = \frac{1}{2}, \\ t_2\,(t) &= s_2\,(t) - \langle s_2, w_1 \rangle w_1\,(t) = t - \frac{1}{2} \cdot 1 = t - \frac{1}{2}, \\ \parallel t_2 \parallel &= \sqrt{\langle t_2, t_2 \rangle} = \sqrt{\int_0^1 \left(t - \frac{1}{2}\right) \cdot \left(t - \frac{1}{2}\right) dt} \\ &= \sqrt{\int_0^1 \left(t^2 - t + \frac{1}{4}\right) dt} = \sqrt{\left(\frac{t^3}{3} - \frac{t^2}{2} + \frac{t}{4}\right)\bigg|_{t=[0,1]}} = \sqrt{\frac{1}{12}}, \\ w_2 &= \frac{t - \frac{1}{2}}{\sqrt{\frac{1}{12}}} = \sqrt{12}t - \frac{\sqrt{12}}{2} = 2\sqrt{3}t - \sqrt{3}. \end{aligned}$$

   It is easy to check that w_2 has unit norm and is orthogonal to w_1.
3. Solve for w_3 :
   Equation:

$$\begin{aligned}
\langle s_3, w_1 \rangle &= \int_0^1 t^2 \cdot 1 dt = \frac{t^3}{3}\bigg|_{t=[0,1]} = \frac{1}{3}, \\
\langle s_3, w_2 \rangle &= \int_0^1 t^2 \cdot \left(2\sqrt{3}t - \sqrt{3}\right) dt = \left(\frac{\sqrt{3}}{2} t^4 - \frac{1}{\sqrt{3}} t^3\right)\bigg|_{t=[0,1]} = \frac{1}{2\sqrt{3}}, \\
t_3 &= s_3 - \langle s_3, w_1 \rangle w_1 - \langle s_3, w_2 \rangle w_2 = t^2 - \frac{1}{3} - \frac{1}{2\sqrt{3}} \cdot \left(2t\sqrt{3} - \sqrt{3}\right) \\
&= t^2 - t - \frac{1}{6}, \\
\parallel t_3 \parallel &= \sqrt{\langle t_3, t_3 \rangle} = \sqrt{\int_0^1 \left(t^2 - t - \frac{1}{6}\right)^2 dt}, \\
w_3\left(t\right) &= \frac{t_3}{\parallel t_3 \parallel}.
\end{aligned}$$

We can check that $[W] = Q$.

Projection Theorem
Introduces the concept of subspace projection and the optimality given by the projection theorem.

Uniqueness of Decompositions

For the next property, we need two quick definitions.

Definition 1 The sum of two spaces $S_1, S_2 \subseteq X$ is the set $S_1 + S_2 = \{x + y : x \in S_1, y \in S_2\}$.

Definition 2 The subspaces $S_1, S_2 \subseteq X$ are disjoint if $S_1 \cap S_2 = \{0\}$.

Theorem 1 Let $V, W \subseteq X$ be subspaces. For each $x \in V + W$ there exists a unique pair $v \in V, w \in W$ such that $x = v + w$ if and only if V and W are disjoint.

Proof: We prove the two directions of this "if and only if" statement separately.

$(\Rightarrow)$ If for each $x \in X$ there exists a unique pair $v, w \in V \times W$ such that $x = v + w$, then the subspaces V and W are disjoint.

Assume there exists a unique pair v, w s.t. $x = v + w$ for each $x \in V + W$. For the sake of contradiction, assume V and W are not disjoint, i.e. there exists some $z \in V \cap W$ such that $z \neq 0$. Pick x and the corresponding v, w such that $x = v + w$. Then,
Equation:

$$\begin{aligned} x &= v + w + z - z, \\ &= v + z + w - z. \end{aligned}$$

We examine these two terms in the equation:
Equation:

$$v + z : v \in V,\ z \in V \Rightarrow v + z \in \text{V},$$

Equation:

$$w - z : w \in W,\ z \in W \Rightarrow w - z \in \text{W}.$$

Also note
Equation:

$$v + z \neq v \quad \text{and} \quad w - z \neq w \quad \text{since} \quad z \neq 0.$$

So we have two pairs of elements from V and W which sum to x, a contradiction to the assumption that these pairs are unique. Therefore, $V \cap W = \{0\}$ and thus V and W are disjoint.

$(\Leftarrow)$ If V and W are disjoint, then for each $x \in V + W$ there exists a unique pair $v, w \in V \times W$ such that $v + w = x$.

To prove uniqueness, we will assume that two distinct pairs exist and show at the end that they are qual to each other. The two pairs we begin with are
Equation:

$$x = v + w,\ v \in V \text{ and } w \in W,$$

Equation:

$$x = p + q,\ p \in V \text{ and } q \in W,$$

Equation:

$$p \neq v \text{ or } q \neq w,$$

so the pairs are distinct from each other. Subtract the equations:
Equation:

$$\begin{aligned} 0 &= p - v + q - w, \\ w - q &= p - v. \end{aligned}$$

We examine these two terms in the equation:
Equation:

$$p - v : p \in V,\ v \in V \Rightarrow p - v \in \text{V},$$

Equation:

$$w - q : w \in W,\ q \in W \Rightarrow w - q \in \text{W}.$$

Since those two terms are equal,
Equation:

$$w - q \in V \quad \text{and} \quad p - v \in \text{W}.$$

which means that $w - q \in V \cap W$ and $p - v \in V \cap W$. Our starting assumption was that $V \cap W = \{0\}$, so therefore $w - q = 0$ and $p - v = 0$ which implies $w = q$ and $p = v$. This statement is a contradiction since we assumed that the pairs were distinct, i.e., $w \neq q$ and $p \neq v$. Therefore, only a unique pair $v \in V$, $w \in W$ exists such that $v + w = x$.

Fact 1 If S is a subspace, then S and $S^{\perp}$ are disjoint. To see this, assume $x \in S \cap S^{\perp}$, which means $x \in S^{\perp}$ and so $\langle x, y \rangle = 0$ for all $y \in S$. Pick $y = x$ so $\langle x, x \rangle = 0 \Rightarrow x = 0$ which means $S \cap S^{\perp} = \{0\}$.

Fact 2 Using Fact 1, we can show that for each $x \in X$ there exists a unique pair $s \in S$ and $t \in S^{\perp}$ such that $x = s + t$. In particular:

- If $x \in S$, $s = x$ and $t = 0$.
- If $x \in S^{\perp}$, $s = 0$ and $t = x$.
- If $x \notin S$ and $x \notin S^{\perp}$, both $s, t \neq 0$.

Projection Theorem

In fact, there is quite a lot more we can say about the projection of a point $x \in X$ into a subspace $S \subseteq X$

Theorem 2 Let X be a Hilbert space and S be a closed subspace of X. For any vector $x \in X$ there exists a unique point $s_0 \in S$ that is closest to x, i.e.,
Equation:

$$s_0 = \min_{s \in S} \| x - s \|$$

In other words, $\| x - s_0 \| \leq \| x - s \|$ for all $s \in S$ with equality only if $s = s_0$.

Furthermore, s_0 is a minimizer of $\| x - s \|$ over $s \in S$ if and only if $x - s_0 \perp S$. In other words, $x - s_0 \in S^{\perp}$.

Proof: To structure our proof, we will restate the theorem into four separate facts:

1. *Existence:* A minimizer of the distance $\| x - s \|$ exists in S.
2. *Orthogonality of the error:* If s_0 is a minimizer, then $x - s_0 \perp S$.
3. *Sufficiency of orthogonality:* If $x - s_0 \perp S$, then s_0 is a minimizer to the distance function.
4. *Uniqueness:* only one point that minimizes the distance exists in S.

Each of these facts is proven separately.

1. Existence: If $x \in S$, then $s_0 = x$ is the minimizer (since 0 is the minimum distance). If $x \notin S$, we denote $\delta = \inf_{s \in S} \| x - s \|$. Note here that we use the infimum, which is the upper bound on all the lower bounds on the distance. The infimum is used because we do not yet know if there exists a point in S that has lowest distance to x. Note also that an infimum exists because norms have a lower bound.

We need to show that for some $s_0 \in S$, we have $\| x - s_0 \| = \delta$. Let $\{s_i\} \subseteq S$ be a sequence that yields $\| x - s_i \| \rightarrow \delta$. We will show that this sequence is a Cauchy sequence; thenm using the fact that the space S is closed and Hilbert, it is complete and therefore the Cauchy sequence must converge to a point $s_0 \in S$.

To prove that the sequence is Cauchy, we use the Parallelogram Law:
Equation:

$$\begin{aligned} \| (s_j - x) + (x - s_i) \|^2 + \| (s_j - x) - (x - s_i) \|^2 &= 2\left(\| s_j - x \|^2 + \| x - s_i \|^2\right), \\ \| s_j - s_i \|^2 &= 2 \| s_j - x \|^2 + 2 \| x - s_i \|^2 - \| s_j + s_i - 2x \|^2, \\ \| s_j - s_i \|^2 &= 2 \| s_j - x \|^2 + 2\| x - s_i \|^2 - 4\| \frac{s_i + s_j}{2} - x \|^2 . \end{aligned}$$

Since S is a subspace, $\frac{s_i + s_j}{2} \in S$ so $\|x - \frac{s_i + s_j}{2}\|^2 \geq \delta^2$, since δ is the infimum. Therefore,
Equation:

$$\| s_j - s_i \|^2 \leq 2 \| x - s_j \|^2 + 2\| s_i - x \|^2 - 4\delta^2 .$$

At this point, we note that $\| s_j - s_i \| \rightarrow 0$ as $i, j \rightarrow \infty$, and so the sequence can be shown to be a Cauchy sequence. Since $\{s_i\}$ is Cauchy, it converges to a point s_0 and by the triangle inequality:
Equation:

$$\| x - s_0 \| \leq \| x - s_j \| + \| s_j - s_0 \| \text{ for each value of } j = 1, 2, ...$$

Now the first term of the inequality goes to δ as $j \rightarrow \infty$ and the second term goes to 0, and so we have that $\| x - s_0 \| \leq \delta + \epsilon$ for all values of $\epsilon > 0$. Since δ was the infimum of the norm on the left hand side, it follows that $\| x - s_0 \| = \delta$, and so we have found that a minimizer exists.

2. Orthogonality of the error: We proceed by contradiction by assuming $x - s_0$ is not orthonormal to S, i.e., that there exists a unit-norm vector $\hat{s} \in S$ such that $\left\langle x - s_0, \hat{s} \right\rangle = \delta \neq 0$. Let $z = s_0 + \delta \hat{s} \in S$. Then

$\parallel x{-}s_0 \parallel \leq \parallel x - z \parallel$. Furthermore,

Equation:

$$\begin{aligned}\parallel x - z \parallel^2 &= \langle x - z, x - z\rangle = \left\langle x - s_0 - \delta\hat{s}, x - s_0 - \delta\hat{s}\right\rangle,\\ &= \langle x - s_0, x - s_0\rangle + \left\langle \delta\hat{s}, \delta\hat{s}\right\rangle - 2\mathrm{Re}\big(\left\langle x - s_0, \delta\hat{s}\right\rangle\big),\\ &= \parallel x{-}s_0 \parallel^2 + |\delta|^2 \parallel \hat{s} \parallel^2 - 2\mathrm{Re}\Big(\bar{\delta}\left\langle x - s_0, \delta\hat{s}\right\rangle\Big).\end{aligned}$$

Now we recognize two simplifications. First, we selected $\hat{s}$ so that $\parallel \hat{s} \parallel^2 = 1$; second, the remaining inner product is equal to $-2\mathrm{Re}\Big(\bar{\delta}\left\langle x - s_0, \delta\hat{s}\right\rangle\Big) = -2\mathrm{Re}\left(\bar{\delta}\delta\right) = -2|\delta|^2$. Therefore,

Equation:

$$\parallel x - z \parallel^2 = \parallel x{-}s_0 \parallel^2 - |\delta|^2 < \parallel x - s_0 \parallel^2 \quad \text{since} \quad \delta \neq 0.$$

This is a contradiction, since we have found a $z \in S$ closer to x than s_0. Thus $\langle x - s_0, s\rangle = 0$ for all $s \in S$, which means that $x - s_0 \perp S$.

3. Sufficiency of orthogonality: Assume $x - s_0 \perp S$. For any $\tilde{s} \in S$ with $\tilde{s} \neq s_0$, we have $\parallel x{-}\tilde{s} \parallel^2 = \parallel x - s_0 + s_0 - \tilde{s} \parallel^2$. We use the Pythagorean theorem (if $a \perp b$ then $\parallel a + b \parallel^2 = \parallel a \parallel^2 + \parallel b \parallel^2$): since $x - s_0 \perp S$ and $s_0 - \tilde{s} \in S$, we have $\parallel x{-}\tilde{s} \parallel^2 = \parallel x{-}s_0 \parallel^2 + \parallel s_0 - \tilde{s} \parallel^2$. Since $s_0 \neq \tilde{s}$, we have $\parallel s_0 - \tilde{s} \parallel > 0$, and so $\parallel x{-}\tilde{s} \parallel^2 > \parallel x - s_0 \parallel^2$. Now, since we picked $\tilde{s} \in S$ arbitrarily, it follows that s_0 is the minimizer of $\parallel x - s \parallel$ over $s \in S$.

4. Uniqueness: We can write $x = x - s_0 + s_0$. From previous parts, we know that $x - s_0 \in S^{\perp}$ and $s_0 \in S$. Since S and $S^{\perp}$ are disjoint, only one pair of values $x - s_0$ and x_0 that holds these two properties exists. Therefore, the minimizer s_0 is unique.

Least Squares Estimation in Hilbert Spaces
Presents leasts squares estimation in subspaces of Hilbert spaces, with applications.

Projections with Orthonormal Bases

Having an orthonormal basis for the subspace of interest significantly simplifies the projection operator.

Lemma 1 Let $x \in X$, a Hilbert space, and let S be a subspace of X. If $\{b_1, b_2, ...\}$ is an orthonormal basis for S, then the closest point $s_0 \in S$ to x is given by $s_0 = \sum_i \langle x, b_i \rangle b_i$.

We begin by noting that
Equation:

$$\sum_i \langle x, b_i \rangle b_i \ \ = \sum_i \langle x - s_0 + s_0, b_i \rangle b_i = \sum_i \langle x - s_0, b_i \rangle b_i + \sum_i \langle s_0, b_i \rangle b_i.$$

Now, since s_0 is the projection of x onto S, we must have that $x - s_0 \perp S$, and so for each basis element b_i we must have $\langle x - s_0, b_i \rangle = 0$. Additionally, since $s_0 \in S$ and $\{b_1, b_2, ...\}$ is an orthonormal basis for S, we must have that $\sum_i \langle s_0, b_i \rangle b_i = s_0$. Thus, we obtain
Equation:

$$\sum_i \langle x, b_i \rangle b_i \ \ = s_0,$$

proving the lemma.

Application: Communications Receiver

Consider the case of a communications receiver that records a continuous-time signal $r(t) = s(t) + n(t)$ over $0 \leq t \leq 1$, where $s(t)$ is one of m codeword signals $\{s_1(t), ..., s_m(t)\}$, and $n(t)$ is additive white Gaussian noise. The receiver must make the best possible decision on the observed codeword given the reading $r(t)$; this usually involves removing as much of the noise as possible from $r(t)$.

We analyze this problem in the context of the Hilbert space $L_2[0,1]$. To remove as much of the noise as possible, we define the subspace $S = \mathrm{span}(\{s_1(t), ..., s_m(t)\})$. Anything that is not contained in this subspace is guaranteed to be part of the noise $n(t)$. Now, to obtain the projection into S, we need to find an orthonormal basis $\{e_1(t), ..., e_n(t)\}$ for S, which can be done for example by applying the Gram-Schmidt procedure on the vectors $\{s_1(t), ..., s_m(t)\}$. The projection is then obtained according to the lemma as
Equation:

$$r_S(t) \ \ = \sum_{i=1}^{n} \langle r(t), e_i(t) \rangle e_i(t),$$

where $\langle r(t), e_i(t)\rangle = \int_0^1 r(t)e_i(t)dt$.

After the projection is obtained, an optimal receiver proceeds by finding the value of k that minimizes the distance

Equation:

$$d_2(r_S(t), s_k(t)) \;= \int_0^1 |r_S(t) - s_k(t)|^2 dt = \int_0^1 r_S(t)^2 dt + \int_0^1 s_k(t)^2 dt - 2\int_0^1 r_S(t)s_k(t)2dt;$$

note here that the first term does not depend on k, so it suffices to find the value of k that minimizes the “cost”

Equation:

$$\begin{aligned} c_k \;&:= \int_0^1 s_k(t)^2 dt - 2\int_0^1 r_S(t)s_k(t)2dt = \langle s_k(t), s_k(t)\rangle - 2\langle r_S(t), s_k(t)\rangle \\ &= \langle s_k(t), s_k(t)\rangle - 2\left\langle \sum_{i=1}^{n} \langle r(t), e_i(t)\rangle e_i(t), s_k(t)\right\rangle \\ &= \langle s_k(t), s_k(t)\rangle - 2\sum_{i=1}^{n} \langle r(t), e_i(t)\rangle \langle e_i(t), s_k(t)\rangle. \end{aligned}$$

In practice, the codeword signals are designed so that their norms $\| \; s_k(t) \; \|_2 = \sqrt{\langle s_k(t), s_k(t)\rangle}$ are all equal. This design choice reduces the problem above to finding the value of k that maximizes the score

Equation:

$$c'_k \;:= \sum_{i=1}^{n} \langle r(t), e_i(t)\rangle \langle e_i(t), s_k(t)\rangle.$$

Thus, the receiver can be designed according to the diagram in [link].

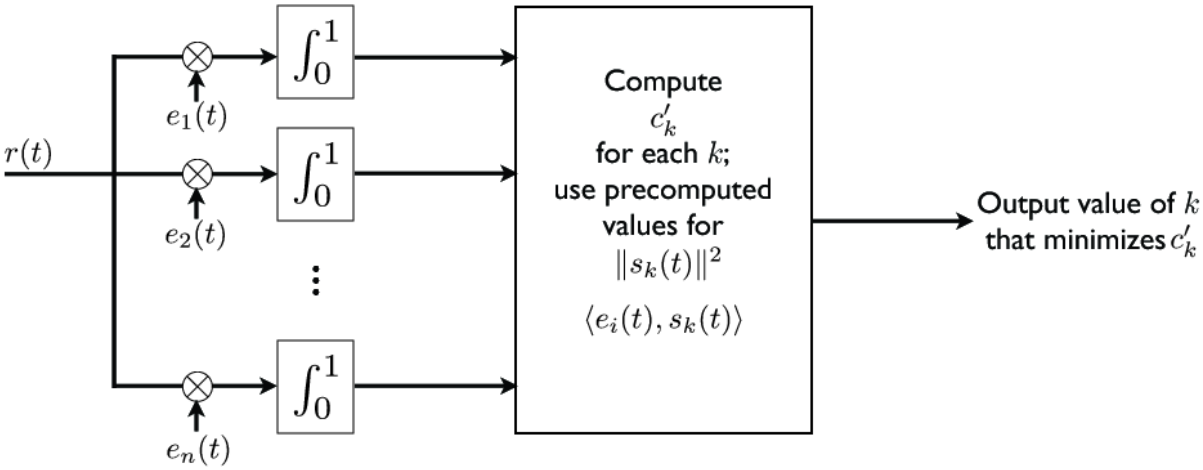

$r(t)$
$e_1(t)$
$e_2(t)$
$e_n(t)$
$\int_0^1$
$\int_0^1$
$\int_0^1$
Compute c'_k for each k;
use precomputed values for
$\|s_k(t)\|^2$
$\langle e_i(t), s_k(t)\rangle$
Output value of k that minimizes c'_k

Diagram of a communications receiver designed in accordance with the projection theorem.

Least Squares Approximation in Hilbert Spaces

Let $y_1, ..., u_n$ be elements of a Hilbert space X and define the closed, finite-dimensional subspace of X given by $S = \text{span}(y_1, ..., y_n)$. We wish to find the best approximation of x in terms of the vectors y_i, that is, the linear combination $\sum_{i=1}^n a_i y_i$ with the smallest error $e = x - \sum_{i=1}^n a_i y_i$. To measure the size of the error, we use the induced norm $\| \ e \ \| = \|x - \sum_{i=1}^n a_i y_i\|$.

To solve this problem, we rely on the projection theorem: we are indeed looking for the closest point to x in $S = \text{span}(y_1, ..., y_n)$. The projection theorem tells us that the closest point $s_0 = \sum_{i=1}^n a_i y_i$ must give $x - s_0 \perp S$, i.e., $e \perp S$, which implies in turn that $\left\langle x - \sum_{i=1}^n a_i y_i, y_j \right\rangle = 0$ for all $j = 1, ..., n$. The requirement can be rewritten as $\langle x, y_j \rangle = \left\langle \sum_{i=1}^n a_i y_i, y_j \right\rangle = \sum_{i=1}^n a_i \langle y_i, y_j \rangle$ for each $j = 1, ..., n$. These requirements can be collected and written in matrix form as

Equation:

$$\underbrace{\begin{bmatrix} \langle x, y_1 \rangle \\ \langle x, y_2 \rangle \\ \vdots \\ \langle x, y_n \rangle \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} \langle y_1, y_1 \rangle & \langle y_2, y_1 \rangle & \cdots & \langle y_n, y_1 \rangle \\ \langle y_1, y_2 \rangle & \langle y_2, y_3 \rangle & \cdots & \langle y_n, y_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle y_1, y_n \rangle & \langle y_2, y_n \rangle & \cdots & \langle y_n, y_n \rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}}_{a}.$$

The coefficients of the best approximation can then be obtained as a vector $a = G^{-1}\beta$, as long as the Gram matrix G is invertible, i.e., it has a nonzero determinant.

In the case that x and y_i are complex-valued vectors, one can rewrite the approximation $\sum_{i=1}^n a_i y_i = Ya$, where a is the coefficient vector denoted above and $Y = [y_1 \ \ldots \ y_n]$ is a matrix that collects the vectors y_i as its columns. The projection theorem requirement then becomes $\langle x - Ma, y_j \rangle = 0$ for $j = 1..., n$, which can be rewritten as $y_j^H (x - Ma) = 0$ and collected as before into the matrix equation

Equation:

$$\begin{aligned} Mj^{H}\left(x - Ma\right) &= 0, \\ M^{H}x - M^{H}Ma &= 0, \\ M^{H}Ma &= M^{H}x, \\ a &= \left(M^{H}M\right)^{-1}M^{H}x = M^{\dagger}x, \end{aligned}$$

which is known as the least squares solution and exists as long as $M^{H}M$ is an invertible matrix. Once these coefficients are obtained, the approximation is equal to $\widehat{x} = Ma = M\left(M^{H}M\right)^{-1}M^{H}x$; therefore, the matrix $P_{M} = M\left(M^{H}M\right)^{-1}M^{H}$ is known as the projection operator.

Application: Channel Equalization

We consider a linear channel with impulse response h that maps an input x into an output y:
Equation:

$$x \to h \to y.$$

We wish to design a linear equalizer of impulse response g for the input x so that after it is run through the equalizer and the channel of impulse response h the output $f \approx x$ (i.e., f is as close as possible to x):
Equation:

$$x \to g \to h \to f.$$

Since the equalizer is linear, the order of g and h can be reversed (this will be discussed in more detail later):
Equation:

$$x \to h \to g \to f.$$

Our design for g will be a finite impulse response filter with tap coefficients g_i; the mapping from input to output index n is therefore given by
Equation:

$$f_n = \sum_{i=1}^{k} g_i y_{n-i}.$$

The error in approximating x_n is given by
Equation:

$$e_n = f_n - x_n = \sum_{i=1}^{k} g_i y_{n-i} - x_n.$$

The total error magnitude over N observations is given by
Equation:

$$E=\sum_{i=0}^{N-1}e_n^2=\sum_{i=0}^{N-1}\sum_{i=1}^{n}(g_iy_{n-i}-x_n)^2.$$

We want to pose this question in terms of error of approximation into a subspace:
Equation:

$$E=\parallel Mg-b\parallel_2^2.$$

By convention, we assume that the values $g_n=0$ and $x_n=0$ for $n<0$ (i.e., $n=0$ is the time of the first observation). It can be easily seen that
Equation:

$$g=\begin{bmatrix}g_1\\ \vdots\\ g_k\end{bmatrix}\text{ and }b=\begin{bmatrix}x_1\\ \vdots\\ x_N\end{bmatrix};$$

formulating M requires a separate study of the sum in [link] for each value of n. For $n=0$,
Equation:

$$\sum_{i=1}^{k}g_iy_{n-i}=\sum_{i=1}^{k}g_iy_{-i}=0,$$

and so terms $n\leq 0$ can be ignored. For $n=1$,
Equation:

$$\sum_{i=1}^{k}g_iy_{n-1}=\sum_{i=1}^{k}g_iy_{1-i}=[y_0,y_{-1},\dots,y_{-k+1}]\begin{bmatrix}g_1\\ \vdots\\ g_k\end{bmatrix}=[y_0,0,\dots,0]\begin{bmatrix}g_1\\ \vdots\\ g_k\end{bmatrix}.$$

Similarly, for $n=2$,
Equation:

$$\sum_{i=1}^{k}g_iy_{n-1}=\sum_{i=1}^{k}g_iy_{2-i}=[y_1,y_0,0,\dots,0]\begin{bmatrix}g_1\\ \vdots\\ g_k\end{bmatrix}.$$

Continuing until $n=N$,
Equation:

$$\sum_{i=1}^{k} g_i y_{n-1} = \sum_{i=1}^{k} g_i y_{N-i} = [y_{N-1}, y_{N-2}, \ldots, y_{N-k}] \begin{bmatrix} g_1 \\ \vdots \\ g_k \end{bmatrix}.$$

The concatenation of these sums as a vector can then be expressed by the matrix-vector product Mg, where the matrix M is given by
Equation:

$$M = \begin{bmatrix} y_0 & 0 & 0 & 0 & \cdots & 0 \\ y_1 & y_0 & 0 & 0 & \cdots & 0 \\ y_2 & y_1 & y_0 & 0 & \cdots & 0 \\ \vdots & & & & & \\ y_{N-1} & y_{N-2} & y_{N-3} & y_{N-4} & \cdots & y_{N-k} \end{bmatrix}.$$

Note that for M to have linearly independent columns (a condition for uniqueness of the solution to g) the number of nonzero values of y_i must be at least k. In this case, the solution
Equation:

$$g = M^{\dagger} b = \left(M^T M\right)^{-1} M^T b$$

minimizes the error as established in the Projection Theorem.

Application: Linear Regression

In linear regression, we are given a set of input/output pairs (x_i, y_i), $i = 1, ..., N$, and we wish to find a linear relationship between inputs and outputs $y_i = a x_i + b$ that minimize the sum of squared errors $E = \sum_{i=1}^{N} \left(y_i - (a x_i + b)^2\right)$. As in previous examples, we seek to pose this minimization problem in terms of the problem considered by the projection theorem: the error $E = \|\ Mg - y\ \|_2^2$, where M is a matrix, y is a vector, and g is the optimization variable vector. One can easily see that the following choice achieves the desired equality:
Equation:

$$M = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}, \; g = \begin{bmatrix} a \\ b \end{bmatrix}, \text{ and } \; y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

As before, the solution that minimizes the error is given by
Equation:

$$g = \left(M^T M\right)^{-1} M^T y,$$

which exists and is unique as long as $G = M^T M$ is invertible, i.e., as long as M has linearly independent columns, i.e., as long as not all x_i are equal. Now, we see that

Equation:

$$\begin{aligned} M^T M &= \begin{bmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^N x_i & N \end{bmatrix}, \\ \left(M^T M\right)^{-1} &= \frac{\begin{bmatrix} N & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \\ M^T y &= \begin{bmatrix} \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}. \end{aligned}$$

Collecting these results, we have that

Equation:

$$g = \frac{\begin{bmatrix} N \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^N x_i\right) \left(\sum_{i=1}^n y_i\right) \\ \left(\sum_{i=1}^N x_i^2\right) \left(\sum_{i=1}^N y_i\right) - \left(\sum_{i=1}^N x_i\right) \left(\sum_{i=1}^N x_i y_i\right) \end{bmatrix}}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2}.$$

Least Squares: Rejoinder

We have studied several examples where an optimization problem can be formulated as

Equation:

$$\hat{g} = \underset{g}{\text{argmin}} \parallel Mg - b \parallel_2^2,$$

where M is a matrix and b and g are column vectors of appropriate sizes.

- When the columns of M are orthonormal (i.e., orthogonal and unit norm), we compute $\hat{g} = M^H b$, i.e., $\hat{g}_i = \langle M_i, b \rangle$, where $\hat{g}_i$ is the i^{th} entry of g and M_i is the i^{th} column of M.
- When the columns of M are linearly independent, we compute $\hat{g} = M^\dagger b = \left(M^H M\right)^{-1} M^H b$. The Moore-Penrose pseudoinverse $M^\dagger = \left(M^H M\right)^{-1} M^H$ is well defined because the Gram matrix $G = M^H M$ of a linearly independent set is invertible.
- When the columns of M are linearly dependent, there exists a vector $a \neq 0$ such that $Ma = 0$. Thus, if a solution g_0 to the optimization exists, then $g_0 + a$ is also a solution: note that $\parallel Mg_0 - b \parallel = \parallel M(g_0 + a) - b \parallel$, and so we lose uniqueness of the minimizer; in fact, we will have infinitely many solutions to the minimization. However, there are ways to rank the solutions and pick a "favorite", e.g., the solution with the smallest norm. This will be considered later in the course.

The Hilbert Space of Random Variables
Describes random variables in terms of Hilbert spaces, defining inner products, norms, and minimum mean square error estimation.

Random Variable Spaces

Probability – Notation Primer

Definition 1 A *random variable* x is defined by a distribution function
Equation:

$$P\left(x\right)={F}_{X}\left(x\right)=\text{Prob}\left(X\le x\right)$$

The *density function* is given by
Equation:

$$\frac{\partial P(x)}{\partial x}={f}_{X}\left(x\right)=\frac{\partial \,\text{Prob}(X\le x)}{\partial x}$$

Definition 2 The *expectation* of a function $g(x)$ over the random variable x is
Equation:

$${E}_{X}\left[g\left(x\right)\right]={\int }_{-\infty }^{\infty }g\left(x\right){f}_{X}\left(x\right)dx$$

Definition 3 Pairs of random variables X,Y are defined by the joint distribution function
Equation:

$$P\left(x,y\right)={F}_{XY}\left(x,y\right)=\text{Prob}\left(X\le x,Y\le y\right)$$

The *joint density function* is given by
Equation:

$$\frac{{\partial }^{2}P\left(x,y\right)}{\partial x\partial y}={f}_{XY}\left(x,y\right)=\frac{{\partial }^{2}\,\text{Prob}\left(X\le x,Y\le y\right)}{\partial x\partial y}$$

The *expectation* of a function $g(x,y)$ is given by

Equation:

$$E_{X,Y}\left[g\left(x,y\right)\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g\left(x,y\right)f_{XY}\left(x,y\right)dx\,dy$$

A Hilbert Space of Random Variables

Definition 4 Let $\{Y_1,\cdots,Y_n\}$ be a collection of zero-mean ($E[Y_i]=0$) random variables. The space H of all random variables that are linear combinations of those n random variables $\{y_1,\cdots,y_n\}$ is a Hilbert space with inner product
Equation:

$$\langle X,Y\rangle=E\left[x\overline{y}\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x\overline{y}f_{XY}\left(x,y\right)\mathrm{d}x\mathrm{d}y.$$

We can easily check that this is a valid inner product:

- $\langle x,x\rangle=E\left[x\overline{x}\right]=\int_{-\infty}^{\infty}\left|x\right|^2f_x\left(x\right)\mathrm{d}x=E\left[\left|x\right|^2\right]\geq 0$;
- $\langle x,x\rangle=0$ if and only if $f_X\left(x\right)=\delta\left(x\right)$, i.e., if X is a random variable that is deterministically zero (and this random variable is the “zero” of this Hilbert space);
- $\langle x,y\rangle=\overline{\langle y,x\rangle}$;
- $\langle x+y,z\rangle=E\left[(x+y)\overline{z}\right]=E\left[x\overline{z}\right]+E\left[y\overline{z}\right]=\langle x,z\rangle+\langle y,z\rangle$;

Note in particular that orthogonality, i.e., $\langle x,y\rangle=0$, implies $E[x\overline{y}]=0$, i.e., x and y are independent random variables. Additionally, the induced norm $\parallel X\parallel=\sqrt{\langle X,X\rangle}=\sqrt{E[|x|^2]}$ is the standard deviation of the zero-mean random variable X.

A Hilbert Space of Random Vectors

One can define random vectors X,Y whose entries are random variables:
Equation:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}.$$

For these, the following inner product is an extension of that given above:
Equation:

$$\langle X, Y \rangle = E\left[y^H x\right] = E\left[\sum_{i=1}^{n} \overline{y_i} x_i\right] = E\left[\text{trace}\left[xy^H\right]\right].$$

The induced norm is
Equation:

$$\| X \| = \sqrt{\langle X, X \rangle} = E\left[\sqrt{x^H x}\right] = E\left[\sqrt{\sum_{i=1}^{N} |x_i|^2}\right],$$

the expected norm of the vector x.

Minimum Mean Square Error Estimation

In an MMSE estimation problem, we consider $Y = AX + N$, where X, Y are two random vectors and N is usually additive white Gaussian noise (Y is $m \times 1$, A is $m \times n$, X is $n \times 1$, and $N \sim \mathcal{N}\left(0, \sigma^2 I\right)$ is $m \times 1$). Due to this noise model, we want an estimate $\widehat{X}$ of X that minimizes $E\left[\| X - \widehat{X} \|^2\right]$; such an estimate has highest likelihood under an additive white Gaussian noise model. For computational simplicity, we often want to restrict the estimator to be linear, i.e.
Equation:

$$\widehat{X} = KY = \begin{bmatrix} K_1^H \\ \vdots \\ K_n^H \end{bmatrix} Y,$$

where K_i^H denotes the i^{th} row of the estimation matrix K and $\widehat{X}_i = K_i^H Y$. We use the definition of the ℓ_2 norm to simplify the equation:
Equation:

$$\min_{K} E\left[\| X - \widehat{X} \|_2^2\right] = \min_{K} E\left[\| X - KY \|_2^2\right] = \min_{K} E\left[\sum_{i=1}^{n} \left(X_i - K_i^H Y\right)^2\right]$$

Since the terms involved in the sum are independent from each other and nonnegative, this minimization can be posed in terms of n individual minimizations: for $i = 1, 2, \ldots, n$, we solve

Equation:

$$\min_{K_i} E\left[\left(X_i - K_i^H Y\right)^2\right] = \min_{K_i} E\left[\left(X_i - \sum_{i=1}^{n} \overline{K_{ij}} Y_j\right)^2\right] = \min_{K_i} \|X_i - \sum_{i=1}^{n} \overline{K_{ij}} Y_j\|,$$

where the norm is the induced norm for the Hilbert space of random variables. Note at this point that the set of random variables $\sum_{i=1}^{n} \overline{K_{ij}} Y_j$ over the choices of K_i can be written as $\mathrm{span}\left(\{Y_j\}_{j=1}^{m}\right)$. Thus, the optimal K_i is given by the coefficients of the closest point in $\mathrm{span}\left(\{Y_j\}_{j=1}^{m}\right)$ to the random variable X_i according to the induced norm for the Hilbert space of random variables. Therefore, we solve for K_i using results from the projection theorem with the corresponding inner product. Recall that given a basis Y_i for the subspace of interest, we obtain the equation $\beta_i = G\left(K_i^H\right)^T = G\overline{K_i}$, where $\beta_{i,j} = \langle X_i, Y_j \rangle$ and G is the Gramian matrix. More specifically, we have

Equation:

$$\underbrace{\begin{bmatrix} \langle X_i, Y_1 \rangle \\ \langle X_i, Y_2 \rangle \\ \vdots \\ \langle X_i, Y_m \rangle \end{bmatrix}}_{\beta_i} = \underbrace{\begin{bmatrix} \langle Y_1, Y_1 \rangle & \langle Y_2, Y_1 \rangle & \cdots & \langle Y_m, Y_1 \rangle \\ \langle Y_1, Y_2 \rangle & \langle Y_2, Y_2 \rangle & \cdots & \langle Y_m, Y_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle Y_1, Y_m \rangle & \langle Y_2, Y_m \rangle & \cdots & \langle Y_m, Y_m \rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} \overline{K_{i1}} \\ \overline{K_{i2}} \\ \vdots \\ \overline{K_{im}} \end{bmatrix}}_{K_i}.$$

Thus, one can solve for $\overline{K_i} = G^{-1}\beta_i$. In the Hilbert space of random variables, we have

Equation:

$$
\begin{aligned}
G &= \begin{bmatrix} E[Y_1Y_1] & E[Y_2Y_1] & \cdots & E[Y_mY_1] \\ E[Y_1Y_2] & E[Y_2Y_2] & \cdots & E[Y_mY_2] \\ \vdots & \vdots & \ddots & \vdots \\ E[Y_1Y_m] & E[Y_2Y_m] & \cdots & E[Y_mY_m] \end{bmatrix} = R_Y, \\
\beta &= \begin{bmatrix} E[X_iY_1] \\ E[X_iY_2] \\ \vdots \\ E[X_iY_m] \end{bmatrix} = \rho_{X_iY}.
\end{aligned}
$$

Here R_Y is the correlation matrix of the random vector Y and ρ_{X_iY} is the cross-correlation vector of the random variable X_i and vector Y. Thus, we have $\overline{K_i} = G^{-1}\beta_i = R_Y^{-1}\rho_{X_iY}$, and so $K_i^H = \rho_{X_iY}^T R_Y^{-1}$. Concatenating all the rows of K together, we get $K = R_{X,Y}R_Y^{-1}$, where $R_{X,Y}$ is the cross-correlation matrix for the random vectors X and Y. We therefore obtain the optimal linear estimator $\widehat{X} = KY = R_{X,Y}R_Y^{-1}Y$.

At first, there may be some confusion on the difference between least squares and minimum mean-square error. To summarize:

- *Least Squares* are applied when the quantities observed are deterministic (i.e., a “single draw” of data or observations).
- *Minimum Mean Square Error Estimation* are applied when random variables are observed under Gaussian noise; one must know a distribution over inputs, and the error must be measured in expectation.

Infinite-Dimensional Hilbert Spaces
Describes the extension of Hilbert spaces to infinite dimensions, including complete orthonormal sequences and Parseval's relation.

While up to now bases have been linked to finite-dimensional spaces and subspaces, it is possible to extend the concept to infinite-dimensional spaces.

Definition 1 Let $(X, R, +, \cdot)$ be a vector space. An infinite sequence of orthonormal vectors $\{e_1, e_2, ...\} \subseteq X$ is said to be a *complete orthonormal sequence* (CONS) in X if for every $x \in X$ there exists a sequence $\alpha_1, \alpha_2, \ldots \in R$ such that $x = \sum_i \alpha_i e_i$.

For the sake of concreteness, an infinite sum is defined as $x = \sum_{i=1}^{\infty} \alpha_i e_i = \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i e_i$. It is easy to see that $\alpha_i = \langle x, e_i \rangle$.

Lemma 1 An orthonormal sequence is complete if and only if the only vector in X orthogonal to each of the e_i's is the zero vector.

Example 1 For the space X of finite-energy complex-valued functions, $R = \mathbb{C}$, a CONS is given by $e_k\left(t\right) = \frac{1}{\sqrt{2\pi}} e^{jkt}$ for $k = 0, \pm 1, \pm 2, \ldots$. These vectors are orthonormal:
Equation:

$$\langle e_k, e_l \rangle = \int_0^{2\pi} \frac{1}{2\pi} e^{jkt} e^{-jlt} dt = \int_0^{2\pi} \frac{1}{2\pi} e^{j(k-l)t} dt = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

The coefficients are given by
Equation:

$$c_k = \langle x, e_k \rangle = \int_0^{2\pi} x\left(t\right) e^{-jkt} dt,$$

and we obtain $x = \sum_k c_k e_k$. This is the sequence behind the Fourier series representation.

Example 2 Let X be the space of bandlimited functions $x(t)$ (i.e., the set of functions with Fourier transform $X(f)$ such that $|X(f)| = 0$ for all $f \notin [-B, B]$). A CONS for this space is given by

Equation:

$$e_k\left(t\right) = \frac{1}{\sqrt{2B}}\text{sinc}\left(2B\left(t - \frac{k}{2B}\right)\right),$$

where $\text{sinc}(t) = (\sin\,(\pi t))/(\pi t)$. It is possible to show that the functions are orthogonal to each other, i.e.,

Equation:

$$\langle e_k, e_l\rangle = \delta_{k,l}\begin{cases}1 & k = l\\ 0 & k \neq l\end{cases}.$$

If x is bandlimited, then it follows that $x\left(t\right) = \sum_k c_k e_k\left(t\right)$, with $c_k = \langle x, e_k\rangle = x\left(k/\left(2B\right)\right)$. The preservation of the norm in the coefficients can also be extended from ONBs to CONS.

Theorem 1 (Completeness Relation) An orthonormal sequence $e_1, e_2, \ldots$ is complete for X if and only if the completeness relation holds for all $x \in X$:

Equation:

$$\parallel x \parallel^2 = \sum_i \ \langle x, e_i\rangle^{\quad 2} = \sum_i |c_i|^2.$$

The sequence $\{e_i\}$ is CONS if and only if

Equation:

$$x = \sum_{i=1}^{\infty}\langle x, e_i\rangle e_i = \lim_{n\to\infty}\sum_{i=1}^{n}\langle x, e_i\rangle e_i = \lim_{n\to\infty} x_n,$$

where $x_n = \sum_{i=1}^{n}\langle x, e_i\rangle e_i$ is the partial sum. We then have $\parallel x \parallel^2 = \parallel x - x_n \parallel^2 + \parallel x_n \parallel^2$ as these two components are orthogonal to each other. Applying a limit on both sides for n, we have

Equation:

$$\parallel x \parallel^2 = \lim_{n\to\infty}\parallel x - x_n \parallel^2 + \lim_{n\to\infty}\parallel x_n \parallel^2 = 0 + \lim_{n\to\infty}\sum_{i=1}^{n}\ \langle x, e_i\rangle^{\quad 2} = \sum_{i=1}^{\infty}|\langle x, e_i\rangle|^2.$$

Theorem 2 Let X be a Hilbert space with a CONS $\{e_1, e_2, \dots\}$. Then for any $x, y \in X$, Parseval's relation holds: $\langle x, y\rangle = \sum_i \langle x, e_i\rangle\langle y, e_i\rangle$.

Using the CONS, we can write the partial sums $x_n = \sum_{i=1}^n \langle x, e_i\rangle e_i$ and $y_n = \sum_{i=1}^n \langle y, e_i\rangle e_i$. We then have

Equation:

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x, y\rangle| &= |\langle x_n, y\rangle - \langle x_n, y\rangle + \langle x_n, y_n\rangle - \langle x, y\rangle| \\ &= |\langle x_n, y_n - y\rangle + \langle x_n - x, y\rangle| \\ &\leq |\langle x_n, y_n - y\rangle| + |\langle x_n - x, y\rangle| \\ &\leq \| x_n \| \| y_n - y \| + \| x_n - x \| \| y_n \| \end{aligned}$$

Letting $n \to \infty$ we have that the upper bound goes to zero, and therefore as $n \to \infty$, we have $\langle x_n, y_n\rangle \to \langle x, y\rangle$. Therefore,

Equation:

$$\begin{aligned} \langle x, y\rangle &= \lim_{n\to\infty} \langle x_n, y_n\rangle = \lim_{n\to\infty} \sum_{i=1}^n \sum_{j=1}^n \langle\langle x, e_i\rangle e_i, \langle y, e_j\rangle e_j\rangle \\ &= \lim_{n\to\infty} \sum_{i=1}^n \sum_{j=1}^n \langle x, e_i\rangle\langle y, e_j\rangle\langle e_i, e_j\rangle = \lim_{n\to\infty} \sum_{i=1}^n \langle x, e_i\rangle\langle y, e_i\rangle \\ &= \sum_{i=1}^\infty \langle x, e_i\rangle\langle y, e_i\rangle. \end{aligned}$$

Linear Functionals

Introduces the definition of a linear functional in a normed space and the norm of a functional.

Many systems in engineering can be characterized as linear systems, taking inputs in one signal space into outputs in another. The most common type of system maps into a space of scalars, defined as maps $f : X \to R$. We will study these maps (known in the mathematics literature as *functionals*) during the next few lectures.

Definition 1 A functional f on X is *linear* if for all $x, y \in X, a, b \in R$, we have that $f(ax + by) = af(x) + bf(y)$.

Example 1 In the case $X = \mathbb{R}^n$, all linear functionals can be written using the form $f\left(x\right) = \sum_{i=1}^{n} a_i x_i$ for some set of scalars $\{a_1, ..., a_n\} \subseteq \mathbb{R}$.

Example 2 For a Hilbert space X, $f\left(x\right) = \langle x, y\rangle$ for any $y \in X$ is a linear functional.

Example 3 For the space $X = C([0, 1])$, the sampling functionals $f\left(x\right) = x\left(t_0\right) = \langle x, \delta(t - t_0)\rangle$ are linear.

Example 4$f(x) =\| x \|$ is not a linear functional (since the triangle inequality is not a strict equality).

Linear functionals are particularly amenable to analysis.

Theorem 1 If a linear functional on a normed space X is continuous at a point $x_0 \in X$, then it is continuous on the entire space X.

Assume $f(x)$ is contiuous at x_0. Let $\{x_n\}$ be a sequence of points converging to x. Then,
Equation:

$$\begin{aligned} |f\left(x_n\right) - f\left(x\right)| &= |f\left(x_n + x_0 - x_0\right) - f\left(x\right)| = |f\left(x_n + x_0\right) - f\left(x_0\right) - f\left(x\right)| \\ &= |f\left(x_n - x + x_0\right) - f\left(x_0\right)|. \end{aligned}$$

It is easy to show that $\{x_n - x + x_0\}$ is a sequence that converges to x_0. Therefore, $f(x_n - x + x_0)$ has to converge to $f(x_0)$ as f is continuous at x_0:
Equation:

$$|f\left(x_n - x + x_0\right) - f\left(x_0\right)\Big| \xrightarrow{n\to\infty} 0.$$

This implies that the sequence $\{f\left(x_n\right)\}$ converges to $f(x)$ and so f is continuous at x. Since x was arbitrary, then f is continuous on X.

Fact 1 For $f(x)$ a linear functional, what is the value of $f(0)$? For any such f,
Equation:

$$f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0.$$

Definition 2 A linear functional f on a normed space X is *bounded* if there exists a constant $M < \infty$ such that $|f(x)| \leq M \parallel x \parallel$ for all $x \in X$. The smallest such element is denoted as the norm of the function $\parallel f \parallel$:
Equation:

$$\parallel f \parallel = \inf_{x \in X} \{M : |f(x)| \leq M \parallel x \parallel, \forall x \in X\}.$$

There are several ways in which we can write this norm:
Equation:

$$\parallel f \parallel = \sup_{x \in X, x \neq 0} \frac{|f(x)|}{\parallel x \parallel} = \sup_{x \in X, \|x\|=1} |f(x)| = \sup_{x \in X, \|x\| \leq 1} |f(x)|$$

Before continuing, we should verify that the defined $\parallel f \parallel$ is a valid norm.

- $\parallel f \parallel = 0 \Leftrightarrow f = z(x) = 0$ for all $x \in X$.
- $\parallel f \parallel \geq 0$: M has to be positive since $|f(x)|$ and $\parallel x \parallel$ are always positive.
- $\parallel \alpha f \parallel = |\alpha| \parallel f \parallel$: can be trivially shown.
- $\parallel f_1 + f_2 \parallel \leq \parallel f_1 \parallel + \parallel f_2 \parallel$: by definition,
  Equation:

$$\begin{aligned} \parallel f_1 + f_2 \parallel & = \sup_{\|x\| \leq 1} \left| (f_1 + f_2)(x) \right| = \sup_{\|x\| \leq 1} \left| f_1(x) + f_2(x) \right| \leq \sup_{\|x\| \leq 1} (|f_1(x)| + |f_2(x)|), \\ & \leq \sup_{\|x\| \leq 1} \left| f_1(x) \right| + \sup_{\|x\| \leq 1} \left| f_2(x) \right| = \parallel f_1 \parallel + \parallel f_2 \parallel. \end{aligned}$$

The following theorem can save us much work in checking for continuity of linear functionals.

Theorem 2 A linear functional on a normed space is bounded if and only if it is continuous.

($\Rightarrow$) Assume f is bounded, and let M be such that $|f(x)| \leq M \parallel x \parallel$ for all $x \in X$. Then for a sequence $\{x_n\} \to 0$ we have $|f(x_n)| \leq M \parallel x_n \parallel$.Therefore, $|f(x)| \to 0 = f(0)$, which implies that f is continuous at $x = 0$. Using Theorem [link], f is continuous on X.

($\Leftarrow$) Assume f is continuous. If we set $\epsilon = 1$ there exist $\delta > 0$ such that if $\parallel x - 0 \parallel < \delta$ then $|f(x) - f(0)| < 1$, i.e.
Equation:

$$\parallel x \parallel < \delta \Rightarrow |f(x)| < 1.$$

Since f is linear, we can write this as
Equation:

$$\parallel x \parallel < 1 \Rightarrow |f\left(x\right)| < \frac{1}{\delta}.$$

So, $\parallel f \parallel \leq \frac{1}{\delta} < \infty$ and f is bounded.

Example 5 Let X be a space of finitely nonzero sequences with $\parallel x \parallel_{\infty} = \max_i |x_i|$. Define a functional f as
Equation:

$$f\left(x\right) = \sum_{i=1}^{\infty} i x_i$$

f is linear because
Equation:

$$\begin{aligned} f(ax+by) &= \sum_{i=1}^{\infty} i(ax+by)_i = \sum_{i=1}^{\infty} i\left(ax_i+by_i\right) = \sum_{i=1}^{\infty}\left(aix_i+biy_i\right) \\ &= a\sum_{i=1}^{\infty} ix_i + b\sum_{i=1}^{\infty} iy_i = af\left(x\right)+bf\left(y\right) \end{aligned}$$

If we want to show that f is unbounded, we must show that for every $M > 0$ there exists $x \in X$ such that $|f(x)| > M \parallel x \parallel$. Let $x = \left(\underbrace{1,1,\ldots,1}_{\lceil M \rceil \text{ times}}, 0, 0, \ldots\right) \in X$, where $\lceil M \rceil$ is the smallest integer greater or equal to M. Then,
Equation:

$$f\left(x\right) = \sum_{i=1}^{\lceil M \rceil} i \cdot 1 \geq \sum_{i=1}^{\lceil M \rceil} 1 = \lceil M \rceil \geq M.$$

Therefore, f is unbounded.

Dual Spaces
Definition of the dual space of a normed space containing all linear operators.

Definition

The set of linear functionals f on X is itself a vector space:

- $(f+g)(x)=f(x)+g(x)$ (closed under addition),
- $(af)(x)=af(x)$ (closed under scalar multiplication),
- $z(x)=0$ is linear and $f+z=f$ (features an addition zero).

Definition 1 The *dual space* (or algebraic dual) X^* of X is the vector space of all linear function on X.

Example 1 If $X=\mathbb{R}^N$ then $f\left(x\right)=\sum_{i=1}^{N}a_ix_i\in X^*$ and any linear functional on X can be written in this form.

Definition 2 Let X be a normed space. The *normed dual*X^* of X is the normed space of all bounded linear functionals with norm $\parallel f\parallel=\text{sup}_{\parallel x\parallel\le 1}\left|f\left(x\right)\right|$.

Since in the sequel we refer almost exclusively to the normed dual, we will abbreviate to “dual” (and ignore the algebraic dual in the process).

Example 2 Let us find the dual of $X=\mathbb{R}^N$. In this space,
Equation:

$$x=\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}\text{ and }\parallel x\parallel=\sqrt{\sum_{i=1}^{n}\left|x_i\right|^2},\text{ where }x_i\in\mathbb{R}.$$

From Example [link], a linear functional f can be written as:
Equation:

$$f\left(x\right)=\sum_{i=1}^{n}a_ix_i=\langle x,a\rangle,\ a\in\mathbb{R}^n$$

There is a one-to-one mapping between the space X and the dual of X:
Equation:

$$f\in\left(\mathbb{R}^n\right)^*\longleftrightarrow a\in\mathbb{R}^n$$

For this reason, $\mathbb{R}^n$ is called self-dual (as there is a one-to-one mapping $\left(\mathbb{R}^n\right)^* \longleftrightarrow \mathbb{R}^n$). Using the Cauchy-Schwartz Inequality, we can show that
Equation:

$$|f(x)| = |\langle x, a\rangle| \leq \| x \| \| a \|,$$

with equality if a is a scalar multiple of x. Therefore, we have that $\| f \| = \| a \|$, i.e., the norms of X and X^* match through the mapping.

Theorem 1 All normed dual spaces are Banach.

Since we have shown that all dual spaces X^* have a valid norm, it remains to be shown that all Cauchy sequences in in X^* are convergent.

Let $\{f_n\}_{n=1}^{\infty} \subseteq X^*$ be a Cauchy sequence in X^*. Thus, we have that for each $\epsilon > 0$ there exists $n_0 \in \mathbb{Z}^+$ such that if $n, m > n_0$ then $\| f_n\left(x\right) - f_m\left(x\right) \| \rightarrow 0$ as $n, m \rightarrow \infty$. We will first show that for an arbitrary input $x \in X$, the sequence $\{f_n\left(x\right)\}_{n=1}^{\infty} \subseteq R$ is Cauchy. Pick an arbitrary $\epsilon > 0$, and obtain the $n_0 \in \mathbb{Z}^+$ from the definition of Cauchy sequence for $\{\overset{*}{f_n}\}_{n=1}^{\infty}$. Then for all $n, m > n_0$ we have that
$|f_n\left(x\right) - f_m\left(x\right)| = |\left(f_n - f_m\right)\left(x\right)| \leq \| f_n\left(x\right) - f_m\left(x\right) \| \cdot \| x \| \leq \epsilon \| x \|$.

Now, since $R = \mathbb{R}$ and $R = \mathbb{C}$ are complete, we have that the sequence $\{f_n\left(x\right)\}_{n=1}^{\infty}$ converges to some scalar $f^*\left(x\right)$. In this way, we can define a new function f^* that collects the limits of all such sequences over all inputs $x \in X$. We conjecture that the sequence $f_n \rightarrow f^*$: we begin by picking some $\epsilon > 0$. Since $\{f_n\left(x\right)\}_{n=1}^{\infty}$ is convergent to $f^*\left(x\right)$ for each $x \in X$, we have that for $n > n_{0,x}$, $\left|f_n\left(x\right) - \overset{*}{f_n}\left(x\right)\right| \leq \epsilon$. Now, let
$\overset{*}{n_0} = \sup_{x \in X, \|x\|=1} n_{0,x}$. Then, for $n > \overset{*}{n_0}$,
Equation:

$$\| f_n - f^* \| = \sup_{x \in X, \|x\|=1} \left|f_n\left(x\right) - f^*\left(x\right)\right| < \epsilon.$$

Therefore, we have found an $\overset{*}{n_0}$ for each ϵ that fits the definition of convergence, and so $f_n \rightarrow f^*$.

Next, we will show that f^* is linear and bounded. To show that f^* is linear, we check that
Equation:

$$f^*\left(ax + by\right) = \lim_{n\rightarrow\infty} f_n\left(ax + by\right) = a \lim_{n\rightarrow\infty} f_n\left(x\right) + b \lim_{n\rightarrow\infty} f_n\left(x\right) = af^*\left(x\right) + bf^*\left(y\right).$$

To show that f^* is bounded, we have that for each $\epsilon > 0$ there exists $n_0^* \in \mathbb{Z}^+$ (shown above) such that if $n > n_0^*$ then

Equation:

$$\left|f^*(x)\right| \leq |f_n(x) - f^*(x)| + |f_n(x)| \leq \epsilon||x|| + ||f_n|| \cdot ||x|| \leq (\epsilon + ||f_n||)||x||.$$

This shows that f^* is linear and bounded and so $f^* \in X^*$. Therefore, $\{f_n^*\}_{n=1}^{\infty}$ converges in X^* and, since the Cauchy sequence was arbitrary, we have that X^* is complete and therefore Banach.

The Dual of ℓ_p

Recall the space $\ell_p = \{(x_1, x_2, ...) : \sum_{i=1}^{\infty} \left|x_i|^p < \infty\right\}$, with the ℓ_p-norm $||x||_p = (\sum_{i=1}^{\infty} \left|x_i|^p\right)^{1/p}$. The dual of ℓ_p is $(\ell_p)^* = \ell_q$, with $q = \frac{p}{p-1}$. That is, every linear bounded functional $f \in (\ell_p)^*$ can be represented in terms of $f(x) = \sum_{n=1}^{\infty} a_n x_n$, where $a = (a_1, a_2, ...) \in \ell_q$. Note that if $p = 2$ then $q = 2$, and so ℓ_2 is self-dual.

Bases for self-dual spaces

Consider the example self-dual space $\mathbb{R}^n$ and pick a basis $\{e_1, e_2, \ldots, e_n\}$ for it. Recall that for each $a \in \mathbb{R}^n$ there exists a bounded linear functional $f_a \in (\mathbb{R}^n)^*$ given by $f_a(x) = \langle x, a \rangle$. Thus, one can build a basis for the dual space $(\mathbb{R}^n)^*$ as $\{f_{e1}, f_{e2}, \ldots, f_{en}\}$.

Riesz Representation Theorem

Since it is easier to conceive a dual space by linking it to elements of a known space (as seen above for self duals), we may ask if there are large classes of spaces who are self-dual.

Theorem 2 (Riesz Representation Theorem) If $f \in H^*$ is a bounded linear functional on a Hilbert space H, there exists a unique vector $y \in H$ such that for all $x \in H$ we have $f(x) = \langle x, y \rangle$. Furthermore, we have $||f|| = ||y||$, and every $y \in H$ defines a unique bounded linear functional $f_y(x) = \langle x, y \rangle$.

Thus, since there is a one-to-one mapping between the dual of the Hilbert space and the Hilbert space ($H^* \approx H$), we say that all Hilbert spaces are self-dual. Pick a linear bounded functional $f \in H^*$ and let $\mathscr{N} \subseteq H$ be the set of all vectors $x \in H$ for which $f(x) = 0$. Note that $\mathscr{N}$ is a closed subspace of H, for if $\{x_n\} \subseteq \mathscr{N}$ is sequence that converges to $x \in H$ then, due to the continuity of f, $f(x_n) \to f(x) = 0$ and so $x \in \mathscr{N}$. We consider two possibilities for the subspace $\mathscr{N}$:

- If $N = H$ then $y = 0$ and $f(x) = \langle x, 0 \rangle = 0$.
- If $N \neq H$ then we can partition $H = \mathscr{N} + \mathscr{N}^{\perp}$ and so there must exist some $z_0 \in \mathscr{N}^{\perp}$, $z_0 \neq 0$. Then, define $z = \dfrac{z_o}{|f(z_o)|}$ to get $z \in \mathscr{N}^{\perp}$ such that $f(z) = 1$.

  Now, pick an arbitrary $x \in H$ and see that
  Equation:

  $$f(x - f(x)z) = f(x) - f(x)f(z) = f(x) - f(x) = 0.$$

  Therefore, we have that $x - f(x)z \in \mathscr{N}$. Since $z \perp \mathscr{N}$, we have that
  Equation:

  $$\begin{aligned} \langle x - f(x)z, z \rangle &= 0, \\ \langle x, z \rangle - f(x)\langle z, z \rangle &= 0, \end{aligned}$$

  and so
  Equation:

  $$f\left(x\right) = \frac{\langle x, z \rangle}{\langle z, z \rangle} = \frac{\langle x, z \rangle}{||z||^2} = \left\langle x, \frac{z}{||z||^2} \right\rangle.$$

  Thus, we have found a vector $y = \frac{z}{||z||^2}$ such that $f(x) = \langle x, y \rangle$ for all $x \in H$. Now, to compute the norm of the functional, consider
  Equation:

  $$|f(x)| = |\langle x, y \rangle| \leq ||x|| \cdot ||y||,$$

  with equality if $x = \alpha y$ for $\alpha \in R$. Thus, we have that $\parallel f \parallel = \parallel y \parallel$.

We now show uniqueness: let us assume that there exists a second $\hat{y} \neq y$, $\hat{y} \in H$, for which $f\left(x\right) = \left\langle x, \hat{y} \right\rangle$ for all $x \in H$. Then we have that for all $x \in H$,
Equation:

$$\begin{aligned} \langle x, y \rangle &= \left\langle x, \hat{y} \right\rangle, \\ \langle x, y \rangle - \left\langle x, \hat{y} \right\rangle &= 0, \\ \left\langle x, y - \hat{y} \right\rangle &= 0, \\ y - \hat{y} &= 0, \\ y &= \hat{y}. \end{aligned}$$

This contradicts the original assumption that $y \neq \hat{y}$. Therefore we have that y is unique, completing the proof.

Linear Functional Extensions

When we consider spaces nested inside one another, we can define pairs of functions that match each other at the overlap.

Definition 3 Let f be a linear functional on a subspace $M \subseteq X$ of a vector space X. A linear functional F is said to be an *extension* of f to N (where N is another subspace of X that satisfies $M \subseteq N \subseteq X$) if F is defined on N and F is identical to f on M.

Here is another reason why we are so interested in bounded linear functionals.

Theorem 3 (Hahn-Banach Theorem) A bounded linear functional f on a subspace $M \subseteq X$ can be extended to a bounded linear functional F on the entire space X with
Equation:

$$\parallel F \parallel = \sup_{x \in X} \frac{|F(x)|}{\parallel x \parallel} = \parallel f \parallel = \sup_{x \in M} \frac{|f(x)|}{\parallel x \parallel} = \sup_{x \in M} \frac{|F(x)|}{\parallel x \parallel}.$$

The proof is in page 111 of Luenberger.

Linear Operators
Introduces linear operators, along with properties and classifications.

We begin our treatment of linear operators, also known as transformations. They can be thought as extensions of functionals that map into arbitrary vector spaces rather than a scalar space.

Definition 1 Let X and Y be be linear vector spaces and $D \subseteq X$. A rule $A : X \to Y$ which associates every element in $x \in D$ with an element $y = A(x) \in Y$ is said to be a *transformation* from $X \to Y$ with domain D.

We have defined D because the transformation may only be defined for some subset of X.

Definition 2 A transformation A :$X \to Y$ where X and Y are vector spaces over a scalar set R, is said to be *linear* if for every $x_1, x_2 \in X$ and all scalars $\alpha_1, \alpha_2 \in R$,
Equation:

$$A\left(\alpha_1 x_1 + \alpha_2 x_2\right) = \alpha_1 A\left(x_1\right) + \alpha_2 A\left(x_2\right)$$

A common type of linear transformation is the transformation $A : \mathbb{R}^n \to \mathbb{R}^m$. In this case, A is an $m \times n$ matrix with real-valued entries (i.e., $A \in \mathbb{R}^{m \times n}$). There are a variety of linear transformations that arise in practice, producing equations of the form $A(x) = y$, with $x \in X$ and $y \in Y$, where X and Y are linear vector spaces. For example, the equation
Equation:

$$\frac{dx}{dt} - ax\left(t\right) = y\left(t\right)$$

may be written in operator notation as $A(x(t)) = y(t)$, where $A : C[T] \to C[T]$ is the operator
Equation:

$$A\left(\cdot\right) = \frac{d(\cdot)}{dt} - a.$$

Often, we will simply write $y = Ax$.

Definition 3 Let $(X, \| \cdot \|_X)$, $(Y, \| \cdot \|_Y)$ be normed vector spaces. A linear operator $A : X \rightarrow Y$ is *bounded* if there exists a constant $M < \infty$ such that $\| Ax - Y| \leq M \| x \|_X$ for all $x \in X$. The smallest M that satisfies this condition is the *norm* of A:

Equation:

$$\| A \|_{X \rightarrow Y} = \max_{x \in X} \frac{\| Ax \|_Y}{\| x \|_X} = \max_{x \in X, \|x\|_X = 1} \| Ax \|_Y = \max_{x \in X, \|x\|_X \leq 1} \| Ax \|_Y.$$

Geometrically, the operator norm $\| A \|$ measures the maximum extent A transforms the unit circle. Thus, $\| A \|$ bounds the amplifying power of the operator A.

Operators possess many properties that are shared with functionals, with similar proofs.

Definition 4 A linear operator $A : X \rightarrow Y$ is continuous on X if it is continuous at any point $x \in X$.

Theorem 1 A linear operator is bounded if and only if it is continuous.

Definition 5 The *sum* of two linear operators $A_1 : X \rightarrow Y$ and $A_2 : X \rightarrow Y$ is defined as $(A_1 + A_2)x = A_1x + A_2x$. Similarly, the *scaling* of a linear operator $A : X \rightarrow Y$ is defined as $(cA)x = c(Ax)$. Both resulting operators are linear as well.

We can also extend the definition of the dual space to operators.

Definition 6 The normed space of all bounded linear operators from $X \rightarrow Y$ is denoted $B(X, Y)$

Are these spaces complete?

Theorem 2 Let X, Y be normed linear spaces. If Y is a complete space then $B(X, Y)$ is complete.

Much of the terminology for operators is drawn from matrices.

Definition 7 Let X be a linear vector space. The operator $I : X \to X$ given by $I(x) = x$ for all $x \in X$ is known as the *identity operator*, and $I \in B(X, Y)$.

Definition 8 Let $A_1 : X \to Y$ and $A_2 : Y \to Z$ be linear operators. The composition of these two operators $(G_2 G_1)x = G_2\left(G_1 x\right)$ is called a *product operator*.

Definition 9 An operator $A : X \to Y$ is *injective* (or *one-to-one*) if for each $y \in Y$ there exists at most one $x \in X$ such that $y = Ax$. In other words, if $Ax_1 = Ax_2$ then $x_1 = x_2$.

Definition 10 An operator $A : X \to Y$ is *surjective* (or *onto*) if for all $y \in Y$ there exists an $x \in X$ such that $y = Ax$.

Definition 11 An operator $A : X \to Y$ is *bijective* if it is injective and surjective.

Lemma 1 If $A_1 : X \to Y$ is a bijective operator, then there exists a transformation $A_2 : Y \to X$. such that $A_2\left(A_1 x\right) = x$ for all $x \in X$.

Note that the lemma above implies $A_2 A_1 = I$. Thus, we say that A_1 is invertible with inverse $A_1^{-1} = A_2$.

Definition 12 An operator $A : X \to X$ is *non-singular* if it has an inverse in $B(X, X)$; otherwise A is *singular*.

In other words, if a transformation $A : X \to X$ is non-singular there exists a transformation A^{-1} such that $AA^{-1} = I$. This extends the concept of singularity from matrices to arbitrary operators.

Properties of Linear Operators
Discusses basic properties and classifications of operators.

We start with a simple but useful property.

Lemma 1 If $A \in B(X, X)$ and $\langle x, Ax \rangle = 0$ for all $x \in X$, then $A = 0$.

Let $a \in \mathbb{C}$, then for any $x, y \in X$, we have that $x + ay \in X$. Therefore, we obtain
Equation:

$$\begin{aligned} 0 &= \langle x + ay, A(x + ay) \rangle, \\ &= \langle x + ay, Ax + aAy \rangle, \\ &= \langle x, Ax \rangle + a\langle x, Ay \rangle + a\langle y, Ax \rangle + |a|^2 \langle y, Ay \rangle. \end{aligned}$$

Since $x, y \in X$, we have that $\langle x, Ax \rangle = \langle y, Ay \rangle = 0$; therefore,
$0 = a\langle x, Ay \rangle + a\langle y, Ax \rangle$.

If we set $a = 1$, then $0 = \langle x, Ay \rangle + \langle y, Ax \rangle$. So in this case, $\langle x, Ay \rangle = -\langle y, Ax \rangle$.

If we set $a = i$, then $0 = -i\langle x, Ay \rangle + i\langle y, Ax \rangle$. So in this case, $\langle x, Ay \rangle = \langle y, Ax \rangle$.

Thus, $\langle x, Ay \rangle = 0$ for all $x, y \in X$, which means $Ay = 0$ for all $y \in X$. So we can come to the conclusion that $A = 0$.

Solutions to Operator Equations

Assume X and Y are two normed linear spaces and $A \in B(X, Y)$ is a bounded linear operator. Now pick $y \in Y$. Then we pose the question: does a solution $\widehat{x} \in X$ to the equation $Ax = y$ exist?

There are three possibilities:

1. A unique solution exists;
2. multiple solutions exist; or
3. no solution exists.

We consider these cases separately below.

1. *Unique solution*: Assume x and x_1 are two solutions to the equation. In this case we have $Ax = Ax_1$. So $A(x - x_1) = 0$. Therefore, $x - x_1 \in \mathscr{N}(A)$. If the solution x is unique then we must have $x = x_1$ and $x - x_1 = 0$. Therefore, $\mathscr{N}(A) = \{0\}$. Since the operator has a trivial null space, then A^{-1} exists. Thus, the solution to the equation is given by $\widehat{x} = A^{-1}y$;

2. *Multiple solutions*: In this case, we may prefer to pick a particular solution. Often, our goal is to find the solution with smallest norm (for example, to reduce power in a communication problem). Additionally, there is a closed-form expression for the minimum-norm solution to the equation $Ax = y$. **Theorem 1** Let X, Y be Hilbert spaces, $y \in Y$, and $A \in B(X, Y)$. Then $\widehat{x}$ is the solution to $Ax = y$ if and only if $\widehat{x} = A^* z$, where $z \in Y$ is the solution of $AA^* z = y$. Let x_1 be a solution to $Ax = y$. Then all other solutions can be written as $\widehat{x} = x_1 - u$, where $u \in \mathscr{N}(A)$. We therefore search for the solution that achieves the minimum value of $||x_1 - u||$ over $u \in \mathscr{N}(A)$, i.e., the closest point to x_1 in $\mathscr{N}(A)$. Assume $\widehat{u}$ is such a point; then, $x_1 - \widehat{u} \perp \mathscr{N}(A)$. Now, recall that $(\mathscr{N}(A))^\perp = \mathscr{R}\left(A^*\right)$, so $\widehat{x} = x_1 - \widehat{u} \in \mathscr{R}\left(A^*\right)$. Thus $\widehat{x} = A^* z$ for some $z \in Y$, and $y = A\widehat{x} = AA^* z$. Note that if AA^* is invertible, then $\widehat{x} = A^* z = A^*\left(AA^*\right)^{-1} y$.
3. *No solution*: In this case, we may aim to find a solution that minimizes the mismatch between the two sides of the equation, i.e., $\| \ y - Ax \ \|$. This is the well-known projection problem of y into $\mathscr{R}(A)$. **Theorem 2** Let X, Y be Hilbert spaces, $y \in Y$, and $A \in B(X, Y)$. The vector $\widehat{x}$ minimizes $||x - Ay||$ if and only if $A^* A\widehat{x} = A^* y$. Denote $u = Ax$ so that $u \in \mathscr{R}(A)$. We need to find the minimum value of $||y - u||$ over $u \in \mathscr{R}(A)$. Assume $\widehat{u}$ is the closest point to y in $\mathscr{R}(A)$, then $y - \widehat{u} \perp \mathscr{R}(A)$, which means $y - \widehat{u} \in (\mathscr{R}(A))^\perp$. Recall that $(\mathscr{R}(A))^\perp = \mathscr{N}\left(A^*\right)$, which implies that $y - \widehat{u} \in \mathscr{N}\left(A^*\right)$. So $A^*\left(y - \widehat{u}\right) = 0$. Thus $A^* y = A^* \widehat{u}$. Now, denoting $\widehat{u} = A\widehat{x}$, we have that $A^* y = A^* A\widehat{x}$. Note that if $A^* A$ is invertible, then we have $\widehat{x} = \left(A^* A\right)^{-1} A^* y$.

Unitary Operator

Definition 1 An operator $A \in B(X, X)$ is said to be *unitary* if $AA^* = A^* A = I$.

This implies that $A^* = A^{-1}$. Unitary operators have norm-preservation properties.

Theorem 3$A \in B(X, X)$ is unitary if and only if $\mathscr{R}(A) = X$ and $||Ax|| = ||x||$ for all $x \in X$.

Let A be unitary, then for any $x \in X$,
Equation:

$$||x|| \quad = \sqrt{\langle x, x \rangle} = \sqrt{\langle x, AA^* x \rangle} = \sqrt{\langle Ax, Ax \rangle} = ||Ax||.$$

Since $I : X \to X$ and $AA^* : \mathscr{R}(A) \to \mathscr{R}(A)$, then $X = \mathscr{R}(A)$. So if $A \in B(X, X)$ is unitary, then $\mathscr{R}(A) = X$ and $||Ax|| = ||x||$ for all $x \in X$. From [link] we can find that
Equation:

$$0 \quad = ||x||^2 - ||Ax||^2 = \langle x, x\rangle - \langle Ax, Ax\rangle = \langle x, x\rangle - \left\langle x, A^* Ax\right\rangle = \left\langle x, x - A^* Ax\right\rangle.$$

Since this is true for all $x \in X$ we have that $x - A^* Ax = 0$ for all $x \in X$, which means that $x = A^* Ax$ for all $x \in X$. Therfore, we must have $A^* A = I$. Additionally, since the operator is unitary, we find that for any $x, y \in X$ we have that $||Ax - Ay|| = ||A(x - y)|| = ||x - y||$. So $x = y$ if and only if $Ax = Ay$, implying that A is one-to-one. Since $\mathscr{R}(A) = X$, then A is onto as well. Thus, A is invertible, which means $A^{-1}A = I$. So $A^{-1}A = I = A^* A$, and therefore $A^{-1} = A^*$. The result is that A is unitary. We have shown that if $\mathscr{R}(A) = X$ and $||Ax|| = ||x||$ for all $x \in X$, then $A \in B(X, X)$ is unitary.

Corollary 4 If X is finite-dimensional, then A is unitary if and only if $||Ax|| = ||x||$ for all $x \in X$.

Fundamental Subspaces of an Operator
Discusses the two fundamental subspaces of a linear operator - range and nullspace - and the rank of an operator.

Definition 1 Let X be a finite dimensional Hilbert Space with orthonormal basis $\{b_1, b_2, \ldots, b_n\}$ and $A : X \to Y$ be a linear operator. The *range* of A is the subspace

Equation:

$$\mathscr{R}\left(A\right) = \{y \in Y : y = Ax \in X\} = [\{Ab_1, Ab_2, ..., Ab_n\}].$$

It is easy to see that $\mathscr{R}(A)$ is a subspace of Y. To show the second equality above, note that if $x \in X$ then it can be written as $x = \sum_{i=1}^{n} a_i b_i$; therefore,

Equation:

$$Ax = A\left(\sum_{i=1}^{n} a_i b_i\right) = \sum_{i=1}^{n} A\left(a_i b_i\right) = \sum_{i=1}^{n} a_i A b_i = \sum_{i=1}^{n} a_i \psi_i,$$

where $\psi_i = Ab_i$. Before we can claim that $\{\psi_i\}$ is a basis for $\mathscr{R}(A)$, we must first show its elements are linearly independent.

Lemma 1 If X be n-dimensional and $A : X \to Y$, then $\mathscr{R}(A)$ has dimension less than for equal to n.

If the set $\{\psi_1, \psi_2, ..., \psi_n\}$ given above is linearly independent then it is a basis for $\mathscr{R}(A)$, and the dimension of $\mathscr{R}(A)$ is n. If they are not linearly independent we must show that $\dim(\mathscr{R}(A)) < n$. If $\{\psi_1, \psi_2, ..., \psi_n\}$ are linearly dependent then there exists a set of scalars $\{\alpha_1, \alpha_2, ..., \alpha_n\}$ such that $\sum_{i=1}^{n} \alpha_i \psi_i = 0$ with at leas one nonzero α_i; we let that be α_1 without loss of generality. We then have

Equation:

$$\psi_1 = \frac{-1}{\alpha_i} \sum_{i=2}^{n} \alpha_i \psi_i,$$

and so $\operatorname{span}\left(\{\psi_1, ..., \psi_n\}\right) = \operatorname{span}\left(\{\psi_2, ..., \psi_n\}\right)$. If the set $\{\psi_2, ..., \psi_n\}$ is linearly independent, then $\dim(\mathscr{R}(A)) = n - 1 \leq n$. Otherwise, iterate this procedure to show that $\dim(\mathscr{R}(A)) < n - 1$.

Definition 2 The *null space* of a A is the set of all points $x \in X$ that map to zero:
Equation:

$$\mathscr{N}(A) = \{x \in X : Ax = 0\}.$$

It is easy to see that $\mathscr{N}(A)$ is a subspace of X.

Lemma 2 An operator A is non-singular if and only if $\mathscr{N}(A) = \{0\}$.

We can extend the concept of rank from matrices to operators.

Definition 3 The *rank* of A is the dimension of $\mathscr{R}(A) = \operatorname{rank}(A)$.

Theorem 1 Let $A : X \to Y$ and X be an n-dimensional space. Then,
Equation:

$$\operatorname{rank}(A) + \dim(\mathscr{N}(A)) = n.$$

Let $\mathscr{N}(A)$ have dimension m. Design an orthonormal basis $e_1, \ldots, e_n$ such that $e_1, \ldots, e_m$ are an orthonormal basis for $\mathscr{N}(A) \subseteq X$. This implies that $\psi_i = Ae_i = 0$, for $i = 1, \ldots, m$. Therefore,
$\mathscr{R}(A) = \operatorname{span}\left(\{\psi_i\}_{i=1}^{n}\right) = \operatorname{span}\left(\{\psi_i\}_{i=m+1}^{n}\right)$, so
$\operatorname{rank}(A) = \dim(\mathscr{R}(A)) \leq n - m$.

Now we need to show that $\{\psi_i\}_{i=m+1}^{n}$ are linearly independent. We use contradiction by assuming that $\{\psi_i\}_{i=m+1}^{n}$ are linearly dependent, which

means that $\sum_{i=m+1}^{n} b_i \psi_i = 0$ for some scalars b_i that are not all zero, i.e.,
Equation:

$$\sum_{i=m+1}^{n} b_i A e_i = A\left(\sum_{i=m+1}^{n} b_i e_i\right) = 0,$$

so we know $\sum_{i=m+1}^{n} b_i e_i \in \mathscr{N}\,(A)$. We also know that $e_i \perp e_j\ , i \neq j$, and so $e_i \perp \mathscr{N}\,(A)$ for $i = m+1, \ldots, n$. This implies in turn that $\sum_{i=m+1}^{n} b_i e_i \perp \mathscr{N}\,(A)$ for all choices of $\{b_i\}$. Because $\sum_{i=m+1}^{n} b_i e_i \in \mathscr{N}\,(A)$, the only possibility is that $\sum_{i=m+1}^{n} b_i e_i = 0$. However, the orthonormal basis vectors e_i for $i = m+1, \ldots, n$ are linear independent, and so we must have that $b_i = 0$. This is a contradiction with our original assumption, implying that the vectors $\{\psi_i\}_{i=m+1}^{n}$ are linearly independent. Therefore, this set of vectors is a basis for $R(A)$ and $\mathrm{rank}(A) = \dim(\mathrm{R}(A)) = n - m$. This in turn implies that $\mathrm{rank}(A) + \dim(\mathscr{N}(A)) = n$.

Adjoint Operators
Definition and properties of an adjoint operator between normed spaces

Adjoint operators allow us to translate inner products in the destination space to inner products in the source space.

Definition 1 Let X and Y be inner product spaces and $A \in B(X, Y)$. The *adjoint operator*A^*: $Y \to X$ is defined by the equation $\langle Ax, y \rangle_Y = \left\langle x, A^* y \right\rangle_X$ for all $x \in X, y \in Y$.

The subindices in the inner product notation clarify the inner product space we refer to, but it is often implicit from the inputs.

Example 1 Pick $X = \mathbb{R}^n$, $Y = \mathbb{R}^n$, and define the operator $A \in B(X, Y)$ by an $m \times n$ matrix; we want to find its adjoint A^*. We appeal to the definition:

$$\langle Ax, y \rangle_Y = y^T\,(Ax) = y^T Ax = \left(y^T A\right)x = \left(A^T y\right)^T x = \left\langle x, A^T y \right\rangle_X$$

so according to the definition, we have that $A^* y = A^T y$, resulting in $A^* = A^T$.

Theorem 1 If $A \in B(X, Y)$ then the adjoint operator $A^* \in B\,(Y, X)$ and $\| A^* \| = \| A \|$.

To see A^* is linear, note that
Equation:

$$\begin{aligned} \left\langle x, A^*\,(a_1 y_1 + a_2 y_2) \right\rangle &= \langle Ax, a_1 y_1 + a_2 y_2 \rangle = \overline{a_1}\,\langle Ax, y_1 \rangle + \overline{a_2}\,\langle Ax, y_2 \rangle = \overline{a_1}\,\left\langle x, A^* y_1 \right\rangle + \overline{a_2}\,\left\langle x, A^* y_2 \right\rangle, \\ &= \left\langle x, a_1 A^* y_1 + a_2 A^* y_2 \right\rangle, \end{aligned}$$

so $A^*\,(a_1 y_1 + a_2 y_2) = a_1 A^* y_1 + a_2 A^* y_2$. We will next show that $\| A^* \| \leq \| A \|$ and $\| A \| \leq \| A^* \|$, which implies that $\| A^* \| = \| A \|$. First, note that
Equation:

$$\begin{aligned} \frac{\| A^* x \|^2}{\| x \|^2} &= \frac{\left\langle A^* x, A^* x \right\rangle}{\| x \|^2} = \frac{\left\langle A A^* x, x \right\rangle}{\| x \|^2} \leq \frac{\| A A^* x \| \| x \|}{\| x \|^2} \leq \frac{\| A \| \| A^* x \|}{\| x \|} \leq \frac{\| A \| \| A^* \| \| x \|}{\| x \|}, \\ &\leq \| A \| \| A^* \|; \end{aligned}$$

since this is true for all $x \in X$, $\| A^* \|^2 \leq \| A \| \| A^* \|$, and so $\| A^* \| \leq \| A \|$.

Next, pick $x_0 \notin \mathscr{N}\,(A)$ and set $y_0 = \frac{Ax_0}{\|Ax_0\|}$. Note that if $\mathscr{N}(A) = X$, then $Ax = 0$ for all x and the adjoint $A^* y = 0$ for all y satisfies the theorem. For such a choice of x_0 and y_0, we note that
Equation:

$$\begin{aligned} \| Ax_0 \|^2 &= \langle Ax_0, Ax_0 \rangle = \langle Ax_0, \| Ax_0 \| y_0 \rangle = \| Ax_0 \| \langle Ax_0, y_0 \rangle = \| Ax_0 \| \left\langle x_0, A^* y_0 \right\rangle, \\ &\leq \| Ax_0 \| \| x_0 \| \| A^* y_0 \| \leq \| Ax_0 \| \| x_0 \| \| A^* \| \| y_0 \| = \| Ax_0 \| \| x_0 \| \| A^* \|, \end{aligned}$$

so $\| Ax_0 \| \leq \| x_0 \| \| A^* \|$, and $\frac{\|Ax_0\|}{\|x_0\|} \leq \| A^* \|$; thus, $\| A \| \leq \| A^* \|$. Therefore, $\| A \| = \| A^* \|$.

Fact 1 Some quick facts on adjoint operators:

1. $I^* = I$.
2. $(A_1 + A_2)^* = A_1^* + A_2^*$.
3. $(\alpha A)^* = \alpha A^*$.
4. $(A_1 A_2)^* = A_2^* A_1^*$.
5. If A^{-1} exists, then $\left(A^{-1}\right)^* = \left(A^*\right)^{-1}$.

Definition 2 An operator $A \in B(X, X)$ is said to be *self-adjoint* if $A^* = A$.

Definition 3 Let X be a Hilbert space. A self-adjoint linear operator $A \in B(X, X)$ is said to be *positive semidefinite* if $\langle x, Ax \rangle \geq 0$ for all $x \in X$.

Example 2 Let $X = L_2\,[0, 1]$ and define an operator $A \in B(X, X)$ by
Equation:

$$y\,(t) = A\,(x\,(t)) = \int_0^1 K\,(t, s) x\,(s)\,ds,\ \ t \in [0, 1],$$

where $\int_0^1 \int_0^1 |K\,(t, s)|^2\,ds\,dt < \infty$. What is its adjoint A^*?

To find the adjoint, we use its definition $\langle Ax, w \rangle = \langle x, A^* w \rangle$:
Equation:

$$\begin{aligned} \langle Ax, w \rangle & = \int_0^1 (Ax)\,(t) w\,(t)\,dt = \int_0^1 \int_0^1 K\,(t, s) x\,(s)\,ds\,w\,(t)\,dt = \int_0^1 \int_0^1 K\,(t, s) x\,(s) w\,(t)\,dt\,ds, \\ & = \int_0^1 x\,(s) \int_0^1 K\,(t, s) w\,(t)\,dt\,ds = \langle x, v \rangle, \end{aligned}$$

where $v\,(s) = \left(A^* w\right)(s) = \int_0^1 K\,(t, s) w\,(t)\,dt$; changing variables, $\left(A^* w\right)(t) = \int_0^1 K\,(s, t) w\,(s)\,ds$.

Example 3 Let $X = L_2\,[0, 1]$ and define an operator $A \in B(X, X)$ by
Equation:

$$y\,(t) = (Ax)\,(t) = \int_0^t K\,(t, s) x\,(s)\,ds.$$

What is its adjoint A^*? Once again, from the definition,
Equation:

$$\begin{aligned} \langle Ax, w \rangle & = \int_0^1 (Ax)\,(t) w\,(t)\,dt = \int_0^1 \int_0^t K\,(t, s) x\,(s)\,ds w\,(t)\,dt = \int_0^1 \int_0^t K\,(t, s) x\,(s) w\,(t)\,ds\,dt, \\ & = \int_0^1 \int_s^1 K\,(t, s) x\,(s) w\,(t)\,dt\,ds = \int_0^1 x\,(s) \int_s^1 K\,(t, s) w\,(t)\,dt\,ds. \end{aligned}$$

We have that $\left(A^* w\right)(s) = \int_s^1 K\,(t, s) w\,(t)\,dt$; changing variables, we obtain the adjoint $\left(A^* w\right)(t) = \int_t^1 K\,(s, t) w\,(s)\,ds$.

Example 4 Let $X = L_2\,[0, 1]$, $Y = \mathbb{R}^N$, and define the operator $A \in B(X, Y)$ as
Equation:

$$y = A\left(x\right) = \begin{bmatrix} x(t_1) \\ x(t_2) \\ \vdots \\ x(t_n) \end{bmatrix}.$$

What is its adjoint A^*? Once again, from the definition of adjoint,
Equation:

$$\left\langle Ax, w \right\rangle_Y = \left\langle x, A^* w \right\rangle_X,$$

where the subscript denotes the corresponding space for the sake of clarity. Then,
Equation:

$$\begin{aligned} \left\langle Ax, w \right\rangle_Y &= \left\langle \left[x\left(t_1\right)\ x\left(t_2\right)\ \ldots\ x\left(t_n\right)\right]^T, w \right\rangle_Y = \sum_{i=1}^{N} x\left(t_i\right) w_i, \\ \left\langle x, A^* w \right\rangle_X &= \int_0^1 x\left(t\right) \left(A^* w\right)\left(t\right) dt = \int_0^1 x\left(t\right) v\left(t\right) dt, \end{aligned}$$

where $v\left(t\right) = \left(A^* w\right)\left(t\right)$. We can successfully match these two inner products by using delta functions to define v. Recall the properties of delta functions: (i) $x\left(t\right)\delta\left(t - t_0\right) = x\left(t_0\right)\delta\left(t - t_0\right)$, (ii) $\int_0^1 \delta\left(t - t_0\right) dt = 1$. Therefore, we may set $v\left(t\right) = \sum_{i=1}^{N} w_i \delta\left(t - t_i\right)$, which then provides
Equation:

$$\begin{aligned} \left\langle x, v \right\rangle_X &= \int_0^1 x\left(t\right) \left(\sum_{i=1}^{N} w_i \delta\left(t - t_i\right) \right) dt = \sum_{i=1}^{N} \left[\int_0^1 x\left(t\right) w_i \delta\left(t - t_i\right) dt) \right] \\ &= \sum_{i=1}^{N} w_i \left[\int_0^1 x\left(t_i\right) \delta\left(t - t_i\right) \right] dt = \sum_{i=1}^{N} w_i x\left(t_i\right) \left(\int_0^1 \delta\left(t - t_i\right) dt \right) = \sum_{i=1}^{N} w_i x\left(t_i\right). \end{aligned}$$

This confirms that $v\left(t\right) = A^* w\left(t\right) = \sum_{i=1}^{N} w_i \delta\left(t - t_i\right)$.

Theorem 2 Let X, Y be normed spaces and $A \in B(X, Y)$, then $\left(\mathscr{R}\left(A\right)\right)^{\perp} = \mathscr{N}\left(A^*\right)$.

We will show the double inclusion $\left(\mathscr{R}\left(A\right)\right)^{\perp} \subseteq \mathscr{N}\left(A^*\right)$ and $\mathscr{N}\left(A^*\right) \subseteq \left(\mathscr{R}\left(A\right)\right)^{\perp}$. Let $y \in \mathscr{N}\left(A^*\right)$, i.e., $A^* y = 0$. Then for all $x \in X$, we have
Equation:

$$\left\langle x, A^* y \right\rangle = 0 = \left\langle Ax, y \right\rangle$$

and so $y \in \left(\mathscr{R}\left(A\right)\right)^{\perp}$, which implies that $\mathscr{N}\left(A^*\right) \subseteq \left(\mathscr{R}\left(A\right)\right)^{\perp}$. Now, pick $y \in \left(\mathscr{R}\left(A\right)\right)^{\perp}$, i.e., for all $x \in X$ we have that $\left\langle Ax, y \right\rangle = 0$. Then, $\left\langle x, A^* y \right\rangle = 0$, and since this is true for all $x \in X$, then $A^* y = 0$ and $y \in \mathscr{N}\left(A^*\right)$. Therefore, $\left(\mathscr{R}\left(A\right)\right)^{\perp} \subseteq \mathscr{N}\left(A^*\right)$. The two inclusions imply that $\left(\mathscr{R}\left(A\right)\right)^{\perp} = \mathscr{N}\left(A^*\right)$. The following statements can be proved in a similar fashion.

Theorem 3 Let X, Y be normed spaces and $A \in B(X, Y)$, then

1. $\overline{\mathscr{R}(A)} = \left[\mathscr{N}\left(A^*\right)\right]^\perp$
2. $\left[\mathscr{R}\left(A^*\right)\right]^\perp = \mathscr{N}\left(A\right)$
3. $\overline{\mathscr{R}\left(A^*\right)} = \left[\mathscr{N}\left(A\right)\right]^\perp$

Example 5 Let $A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \\ 2 & 3 \end{bmatrix}$, then $\mathscr{R}\left(A\right) = \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right\}$, $A^* = \begin{bmatrix} 2 & 0 & 2 \\ 1 & 2 & 3 \end{bmatrix}$, and $\mathscr{N}\left(A^*\right) = \text{span}\left\{ \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\}$. It is easy to check that $\left(\mathscr{R}\left(A\right)\right)^\perp = \mathscr{N}\left(A^*\right)$.

Matrix Representations of Linear Operators
Introduces matrix representations of linear operators, with examples.

Linear operators involving finite-dimensional spaces can be represented in terms of matrices. Assume that X and Y are finite-dimensional spaces and $A \in B(X, Y)$. Let I$\{e_i\}$ be a orthonormal basis for X so that for all $x \in X$ we have $x = \sum_i \langle x, e_i\rangle e_i$, giving x the unique set of coefficients $a_i = \langle x, e_i\rangle$. Similarly, let $\{\tilde{e}_i\}$ be an orthonormal basis for Y so that for $y \in Y$ we have $y = \sum_i \langle y, \tilde{e}_i\rangle \tilde{e}_i$, giving y the unique set of coefficients $b_i = \langle y, \tilde{e}_i\rangle$. We will now show that the map $x \to y = A(x)$ can be represented in terms of their coefficient vectors as $a \to b = \widetilde{A}a$, where $\widetilde{A}$ is a matrix.

Recall that $\psi_i = Ae_i \in Y$, so it can be written as $\psi_i = \sum_j \langle \psi_i, \tilde{e}_j\rangle \tilde{e}_j$. Therefore,
Equation:

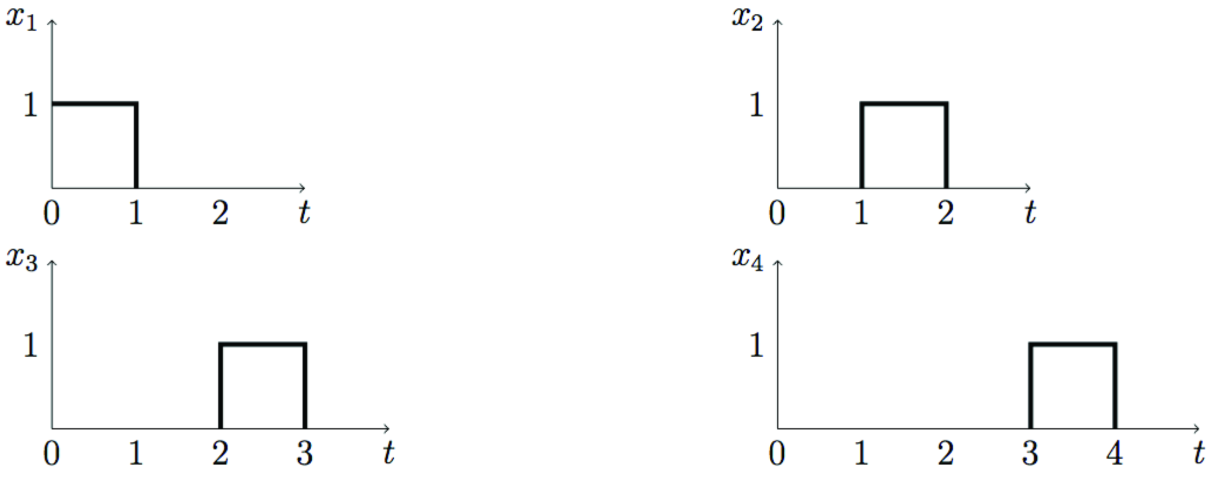
$$\begin{aligned} y &= Ax = A\left(\sum_i \langle x, e_i\rangle e_i\right) = \sum_i \langle x, e_i\rangle Ae_i = \sum_i \langle x, e_i\rangle \psi_i = \sum_i \sum_j \langle x, e_i\rangle \langle \psi_i, \tilde{e}_j\rangle \tilde{e}_j, \\ &= \sum_i \left(\sum_j \langle x, e_i\rangle \langle \psi_i, \tilde{e}_j\rangle\right) \tilde{e}_j. \end{aligned}$$

Due to the uniqueness of coefficients for y in $\{\tilde{e}_j\}$, we have that for each j,
Equation:

$$\begin{aligned} \sum_i \underbrace{\langle x, e_i\rangle}_{a_i} \langle \psi_i, \tilde{e}_j\rangle &= \underbrace{\langle y, \tilde{e}_j\rangle}_{b_j}, \\ \sum_i \underbrace{\langle \psi_i, \tilde{e}_j\rangle}_{\widetilde{A}_{j,i}} a_i &= b_j. \end{aligned}$$

So we have found a matrix $\widetilde{A}$ with entries $\widetilde{A}_{j,i} = \langle \psi_i, \tilde{e}_j\rangle$ that provides $\widetilde{A}a = b$. Note that the matrix will be of size $\dim(Y) \times \dim(X)$.

Example 1 Consider the space $X \subseteq L_2\,[0, 1]$ defined by $X = \mathrm{span}(\{x_1, x_2, x_3, x_4\})$, given below:



Functions in Example 1.

and the space $Y \subseteq L_2\left[0,4\right]$ given by $Y = \text{span}(\{y_1, y_2\})$, where $y_1\left(t\right) = \sqrt{2}\,\cos\left(2\pi t\right)$ and $y_2\left(t\right) = \sqrt{2}\,\cos\left(4\pi t\right)$. We define an operator $A : X \to Y$ as

Equation:

$$y\left(t\right) = A\left(x\left(t\right)\right) = \left(\int_0^3 x\left(t\right)dt\right)\,\cos\left(2\pi t\right) + \left(\int_1^4 x\left(t\right)dt\right)\,\cos\left(4\pi t\right).$$

It is easy to see that an orthonormal basis for X is given by the functions $e_i\left(t\right) = x_i\left(t\right)$. One can also show that an orthonormal basis for Y is given by the functions $\tilde{e}_1\left(t\right) = \frac{1}{\sqrt{2}}\,\cos\left(2\pi t\right)$ and $\tilde{e}_2\left(t\right) = \frac{1}{\sqrt{2}}\,\cos\left(4\pi t\right)$. For this choice of orthonormal bases for X and Y, the transformed basis elements from X are given by

Equation:

$$\begin{aligned}
\psi_1\left(t\right) &= A\left(x_1\left(t\right)\right) = \cos\left(2\pi t\right) = \sqrt{2}\tilde{e}_1,\\
\psi_2\left(t\right) &= A\left(x_2\left(t\right)\right) = \cos\left(2\pi t\right) + \cos\left(4\pi t\right) = \sqrt{2}\left(\tilde{e}_1 + \tilde{e}_2\right),\\
\psi_3\left(t\right) &= A\left(x_3\left(t\right)\right) = \cos\left(2\pi t\right) + \cos\left(4\pi t\right) = \sqrt{2}\left(\tilde{e}_1 + \tilde{e}_2\right),\\
\psi_4\left(t\right) &= A\left(x_4\left(t\right)\right) = \cos\left(4\pi t\right) = \sqrt{2}\tilde{e}_2.
\end{aligned}$$

It is then easy to check that the entries of the matrix are given by

$\widetilde{A}_{1,1} = \langle \psi_1, \tilde{e}_1 \rangle = \sqrt{2},$	$\widetilde{A}_{1,2} = \langle \psi_2, \tilde{e}_1 \rangle = \sqrt{2},$	$\widetilde{A}_{1,3} = \langle \psi_3, \tilde{e}_1 \rangle = \sqrt{2},$	$\widetilde{A}_{1,4} = \langle \psi_4, \tilde{e}_1 \rangle = 0,$
$\widetilde{A}_{2,1} = \langle \psi_1, \tilde{e}_2 \rangle = 0,$	$\widetilde{A}_{2,2} = \langle \psi_2, \tilde{e}_2 \rangle = \sqrt{2},$	$\widetilde{A}_{2,3} = \langle \psi_3, \tilde{e}_2 \rangle = \sqrt{2},$	$\widetilde{A}_{2,4} = \langle \psi_4, \tilde{e}_2 \rangle = \sqrt{2}.$

Thus, the matrix representation for the operator A using these orthonormal bases is given by

Equation:

$$\widetilde{A} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} & 0 \\ 0 & \sqrt{2} & \sqrt{2} & \sqrt{2} \end{bmatrix}.$$

Eigendecomposition of Linear Operators
Introduces the concept of the eigendecomposition of a linear operator, with properties and examples.

Definition 1 A scalar λ is an *eigenvalue* of $A \in B(X, X)$ if there exists a vector e (dubbed the **eigenvector** for λ) such that $Ae = \lambda e$.

Multiples of eigenvectors are also eigenvectors, as shown below:
Equation:

$$\begin{aligned} A(ce) &= c(Ae) \\ &= c\lambda e \\ &= \lambda(ce) \end{aligned}$$

Definition 2 The *eigenspace* of A corresponding to λ is defined by $\epsilon_\lambda = \{e \in X : Ae = \lambda e\}$.

For example, if a given eigenvalue λ has two eigenvectors, the eigenspace is given by $[\{e_1, e_2\}]$ where $Ae_1 = \lambda e_1$ and $Ae_2 = \lambda e_2$.

Definition 3 An operator $A \in B(X, X)$ is said to be *self-adjoint* if $A^* = A$, i.e., $\langle Ax, y \rangle = \langle x, Ay \rangle$ for all $x, y \in X$.

If $X = \mathbb{R}^N$, a self-adjoint operator corresponds to a symmetric matrix.

Theorem 1 All eigenvalues of a self-adjoint operator are real.

Let λ be a complex eigenvalue of a self-adjoint operator A. Then $Ae = \lambda e$ for some e. For such an e, we have
Equation:

$$\lambda \| \ e \ \|^2 = \lambda \langle e, e \rangle = \langle \lambda e, e \rangle = \langle Ae, e \rangle = \left\langle e, A^* e \right\rangle = \langle e, Ae \rangle = \langle e, \lambda e \rangle = \overline{\lambda} \langle e, e \rangle.$$

Therefore we know that lambda is the same as its complex conjugate ($\lambda = \overline{\lambda}$). The only way for this to be possible is if the imaginary part of λ is zero, and therefore $\lambda \in \mathbb{R}$.

Theorem 2 If λ_1, λ_2 are distinct eigenvalues of a self-adjoint operator $A \in B(X, X)$, then $\epsilon_{\lambda_1} \perp \epsilon_{\lambda_2}$.

Assume we pick some arbitrary $e_1 \in \epsilon_{\lambda_1}$ and $e_2 \in \epsilon_{\lambda_2}$. Then,
Equation:

$$\lambda_1 \langle e_1, e_2 \rangle = \langle \lambda_1 e_1, e_2 \rangle = \langle Ae_1, e_2 \rangle = \langle e_1, Ae_2 \rangle = \langle e_1, \lambda_2 e_2 \rangle = \lambda_2 \langle e_1, e_2 \rangle.$$

Therefore we know that $\lambda_1 \langle e_1, e_2 \rangle = \lambda_2 \langle e_1, e_2 \rangle$. Since we chose two distinct eigenvalues, we know that $\lambda_1 \neq \lambda_2$. Therefore we must have $\langle e_1, e_2 \rangle = 0$. This implies $e_1 \perp e_2$. Since e_1 and e_2 were chosen arbitrarily, this implies $\epsilon_{\lambda_1} \perp \epsilon_{\lambda_2}$.

Theorem 3 (Schur's Lemma) Let X be an N-dimensional Hilbert space and $A \in B(X, X)$. Then there exists an orthonormal basis $\{\varphi_1, \varphi_2, ..., \varphi_N\}$ for X and a set of coefficients $\{A_{ij}\}_{i,j=1}^{N}$ such that $A\varphi_j = \sum_{i=1}^{j} A_{ij}\varphi_i$ for $j = 1, 2, \ldots N$.

An important note: since $\{\varphi_j\}$ makes an orthonormal basis for the domain X and range X, there exist coefficients a_i such that $A\varphi_j = \sum_{i=1}^{N} a_i\varphi_i$ for $j = 1, 2, \ldots. N$.

Example 1 Recall the matrix representation of an operator: given $A \in B(X, X)$ and an orthonormal basis $\{\varphi_i\}_{i=1}^{N}$ for X, we can write the matrix representation $\widetilde{A}$ of the operator A with entries

Equation:

$$\widetilde{A}_{ij} = \langle A\varphi_j, \varphi_i \rangle = \left\langle \sum_{k=1}^{j} A_{kj}\varphi_k, \varphi_i \right\rangle = \sum_{k=1}^{j} A_{kj} \langle \varphi_k, \varphi_i \rangle = \begin{cases} A_{ij}, & \text{if } i \leq j, \\ 0, & \text{if } i > j. \end{cases}$$

We can represent $\widetilde{A}$ as an upper-triangular matrix (where x represents a non-zero entry in the matrix):

Equation:

$$\widetilde{A} = \begin{bmatrix} x & x & x & \cdots & x \\ 0 & x & x & \cdots & x \\ 0 & 0 & x & \cdots & x \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & x \end{bmatrix}$$

This shows that there exists an orthonormal basis φ for which the matrix representation of A is an upper-triangular matrix.

Theorem 4 If X is an N-dimensional space and $A \in B(X, X)$, then there exists an orthonormal basis $\{\varphi_i\}_{i=1}^{N}$ such that $A\varphi_i = \lambda_i\varphi_i$ for a set of scalars $\lambda_1, \lambda_2, \ldots \lambda_N$. The matrix representation $\widetilde{A}$ is a diagonal matrix whose entries are the eigenvalues; that is, $\widetilde{A} = \text{diag}\left(\{\lambda_i\}\right)$.

Note that, according to the theorem, we can fully represent the operator A by the aforementioned orthonormal basis $\{\varphi_i\}$ and the diagonal matrix $\widetilde{A}$.

Pick the orthonormal basis $\{\varphi_1, \varphi_2, \ldots \varphi_N\}$ specified by Schur's Lemma. For $i < j$, we have:

Equation:

$$A_{ij} = \langle A\varphi_j, \varphi_i \rangle = \left\langle \sum_{k=1}^{j} A_{kj}\varphi_k, \varphi_i \right\rangle = \langle \varphi_j, A\varphi_i \rangle = \left\langle \varphi_j, \sum_{k=1}^{i} A_{ki}\varphi_k \right\rangle = \sum_{k=1}^{i} \widetilde{A_{ki}} \langle \varphi_j, \varphi_k \rangle.$$

For $k < j$, the term in the sum is equal to zero. We then have:
Equation:

$$A\varphi_j = \sum_{k=1}^{j} A_{kj}\varphi_k = A_{jj}\varphi_j.$$

Thus, the only non-zero entries of the representation matrix $\widetilde{A}$ are the diagonal entries. Furthermore, these entries are eigenvalues of A.

Recall that to compute Ax using its matrix representation $\widetilde{A}$, there are three steps:

1. Represent x using the orthonormal basis $\{\varphi_i\}$ and collect the coefficients into a vector c.
2. Perform the matrix-vector product $d = \widetilde{A}c$.
3. Then obtain $Ax = b = \sum_{i=1}^{N} d_i\varphi_i$.

A matrix-vector product with a dense matrix is computationally intensive. If we can diagonalize A, we can find the matrix-vector product Ax by performing the operation $\widetilde{A}c$, where $\widetilde{A}$ is a diagonal matrix, making the operation more efficient.

The Karhuenen-Loeve Transform
Introduces the Karhuenen-Loeve transform, with applications.

Define the random vector
Equation:

$$X=\begin{bmatrix}x_1\\x_2\\\vdots\\x_N\end{bmatrix}$$

with mean zero and covariance matrix $R_X=E\left[XX^*\right]$; this matrix is symmetric and positive semidefinite.

Lemma 1 Every eigenvalue of R_X is real and non-negative.

Let e be an eigenvector of R_X with eigenvalue λ.
Equation:

$$\lambda\|\ e\ \|^2=\lambda\left\langle e,e\right\rangle=\left\langle\lambda e,e\right\rangle=\left\langle Ae,e\right\rangle\geq 0.$$

The last statement falls out by the definite of positive semi-definite. We have $\lambda\|\ e\ \|^2\geq 0$. Since $\|\ e\ \|^2\geq 0$, it follows that $\lambda\geq 0$, i.e. all the eigenvalues are non-negative. The eigenvectors of the matrix R_X provide an orthonormal basis $\{\varphi_1,\varphi_2,\ldots\varphi_N\}$, which can be collected into an orthonormal basis matrix $\varphi=[\varphi_1\ \varphi_2\ ...\ \varphi_N]$. Then let $y=\varphi^*x$. We have:
Equation:

$$R_Y=E\left[yy^*\right]=E\left[\varphi^*xx^*\varphi\right]=\varphi^*E\left[xx^*\right]\varphi=\varphi^*R_X\varphi.$$

Let us look at the adjoint of φ^*R_X:
Equation:

$$\left(\varphi^*R_X\right)^*=R_X\varphi=R_X\left[\varphi_1\ \varphi_2\ ...\ \varphi_N\right]=\left[\lambda_1\varphi_1\ \lambda_2\varphi_2\ ...\ \lambda_N\varphi_N\right].$$

If we take the adjoint again, we get
Equation:

$$\varphi^* R_X = \begin{bmatrix} \varphi_1^* \lambda_1 \\ \varphi_2^* \lambda_2 \\ \vdots \\ \varphi_N^* \lambda_N \end{bmatrix}.$$

Going back to our derivation of R_Y:
Equation:

$$\begin{aligned} R_Y &= \varphi^* R_X \varphi = \begin{bmatrix} \varphi_1^* \lambda_1 \\ \vdots \\ \varphi_N^* \lambda_N \end{bmatrix} [\varphi_1 \quad \cdots \quad \varphi_N] \\ &= \begin{bmatrix} \lambda_1 \langle \varphi_1, \varphi_1 \rangle & \lambda_1 \langle \varphi_1, \varphi_2 \rangle & \cdots & \lambda_1 \langle \varphi_1, \varphi_N \rangle \\ \lambda_2 \langle \varphi_2, \varphi_1 \rangle & \lambda_2 \langle \varphi_2, \varphi_2 \rangle & \cdots & \lambda_2 \langle \varphi_2, \varphi_N \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_N \langle \varphi_N, \varphi_1 \rangle & \lambda_N \langle \varphi_N, \varphi_2 \rangle & \cdots & \lambda_N \langle \varphi_N, \varphi_N \rangle \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \end{aligned}$$

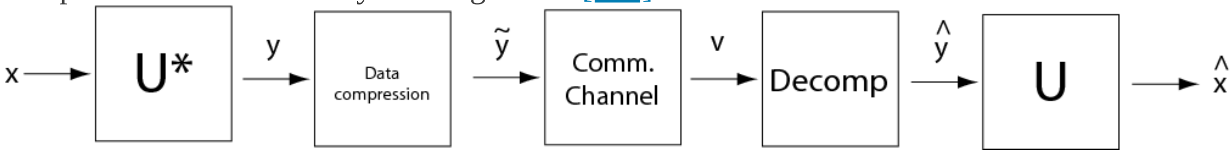
The matrix φ is known as the KLT matrix defined by R_X. The transformation given by the KLT matrix provides a set of random variables $y_i = \langle \varphi_i, x \rangle$ that are uncorrelated.

Example 1 (Whitening Filter) For a random vector X, R_X has positive eigenvalues. Let us write $R_Y^{-1/2} = \text{diag}\left(\lambda_1^{-1/2}, \ldots \lambda_N^{-1/2}\right)$ and $z = R_Y^{-1/2} y$ where $y = \varphi^* x$. We have
Equation:

$$R_Z \; = E\left[zz^*\right] = E\left[R_Y^{-1/2} yy^* R_Y^{-1/2}\right] = R_Y^{-1/2} E\left[yy^*\right] R_Y^{-1/2} = R_Y^{-1/2} R_Y R_Y^{-1/2} = I.$$

The matrix $R_Y^{-1/2} \varphi^*$ is known as a "whitening filter", as it maps an arbitrary random vector x to a “white Gaussian noise” vector z.

Example 2 (Transform Coding) Let $U : \mathbb{C}^n \to \mathbb{C}^n$ is a unitary operator. Assume we have a signal $x \in \mathbb{C}^n$ that we want to send it through a channel by only sending k numbers or "items", where $k < n$; in words, we wish to compress the signal x. The block diagram for the compression/transmission system is given in [link].



Block diagram for a transform coding system.

We want to minimize $E\left[\| x - \widehat{x} \|\right]$ given k by choosing the optimal transformation U. We know $y = U^* x$ which implies $x = Uy$ since U is unitary. Therefore,
Equation:

$$\| x - \widehat{x} \| = \| Uy - U\hat{y} \| = \| U \left(y - \hat{y}\right) \| = \| y - \hat{y} \|.$$

This means that we can minimize $\| y - \hat{y} \|$ in place of $\| x - \widehat{x} \|$. For simplicity, we choose a basic means of compression that preserves only the first k entries of y:
Equation:

$$\widehat{y_i} = \begin{cases} y_i, & \text{if } i = 1, 2, \dots k, \\ 0, & \text{if } i = k+1, k+2, \dots n. \end{cases}$$

We then have $E\left[\| x - \widehat{x} \|^2\right] = E\left[\| y - \hat{y} \|^2\right] = E[\sum_{i=k+1}^{n} \left|y_i\right|^2] = \sum_{i=k+1}^{n} E[|y_i|^2]$. Therefore,
Equation:

$$\begin{aligned} \min_{\widehat{x}} \left(E[\| x - \widehat{x} \|^2]\right) &= \min_U \left(\sum_{i=k+1}^{n} E[|y_i|^2]\right) = \min_U \left(\sum_{i=k+1}^{n} E[| \langle x, u_i \rangle |^2]\right), \\ &= \min_U \left(\sum_{i=k+1}^{n} E\left[u_i^T x x^T u_i\right]\right) = \min_U \left(\sum_{i=k+1}^{n} u_i^T E\left[x x^T\right] u_i\right), \\ &= \min_U \left(\sum_{i=k+1}^{n} u_i^T R_X u_i\right). \end{aligned}$$

It turns out that the choice of transform basis U that minimizes this amount is provided by the eigendecomposition of R_X, as specified by the following theorem.

Theorem 1 Let X be a length-n random vector with covariance matrix $R_X = E\left[x x^*\right]$ that has eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 ... \geq \lambda_n \geq 0$ and matching eigenvectors $\varphi_1, \varphi_2, ... \varphi_N$. Let x_M be the orthogonal projection of x onto a subspace M of dimension k. Then
Equation:

$$E[||x - x_M||^2] \geq \sum_{i=k+1}^{M} \lambda_i,$$

with equality if $M = \text{span}(\{\varphi_1, \varphi_2, ...\varphi_k\})$.

From equation [link], we have

Equation:

$$\min_M E[||x - x_M||^2] = \min_U \left(\sum_{i=k+1}^{n} u_i^T R_X u_i \right) = \min_U \left(\sum_{i=k+1}^{n} u_i^T \varPhi \varLambda \varPhi^T u_i \right),$$

where $R_X = \varPhi \varLambda \varPhi^T$ is the eigendecomposition of R_X. Now, since

Equation:

$$\varPhi^T u_i = \begin{bmatrix} \langle u_i, \varphi_1 \rangle \\ \langle u_i, \varphi_2 \rangle \\ \vdots \\ \langle u_i, \varphi_n \rangle \end{bmatrix} \text{ and } \varLambda \varPhi^T u_i = \begin{bmatrix} \langle u_i, \varphi_1 \rangle \lambda_1 \\ \langle u_i, \varphi_2 \rangle \lambda_2 \\ \vdots \\ \langle u_i, \varphi_n \rangle \lambda_n \end{bmatrix},$$

we have that $u_i^T \varPhi \varLambda \varPhi^T u_i = \left(\varPhi^T u_i \right)^T \varLambda \varPhi^T u_i = \sum_{i=j}^{n} |\langle u_i, \varphi_j \rangle|^2 \lambda_j$. Plugging this into [link], we have

Equation:

$$\min_M E[||x - x_M||^2] = \min_U \left(\sum_{i=k+1}^{n} \sum_{i=j}^{n} |\langle u_i, \varphi_j \rangle|^2 \lambda_j \right) = \min_U \left(\sum_{i=j}^{n} \lambda_j \sum_{i=k+1}^{n} |\langle u_i, \varphi_j \rangle|^2 \right).$$

Now, denote $\alpha_j = \sum_{i=k+1}^{n} |\langle u_i, \varphi_j \rangle|^2$, and see that

Equation:

$$\sum_{j=1}^{n} \alpha_j = \sum_{j=1}^{n} \sum_{i=k+1}^{n} \left| \langle u_i, \varphi_j \rangle \right|^2 = \sum_{i=k+1}^{n} \sum_{j=1}^{n} \left| \langle u_i, \varphi_j \rangle \right|^2 = \sum_{i=k+1}^{n} ||u_i||^2 = n - k,$$

as all u_i are unit-norm. Now, we have that

Equation:

$$\min_M E[||x - x_M||^2] = \min_U \left(\sum_{i=j}^{n} \lambda_j \alpha_j \right) = \min_U \left(\sum_{j=1}^{k} \lambda_j \alpha_j - \sum_{j=k+1}^{n} \lambda_j (1 - \alpha_j) + \sum_{j=k+1}^{n} \lambda_j \right).$$

Since the λ_k are monotonically decreasing, we have that

Equation:

$$\begin{aligned}
\min_M E\left[||x-x_M||^2\right] &\geq \min_U \left(\sum_{j=1}^{k} \lambda_k \alpha_j - \sum_{j=k+1}^{n} \lambda_k \left(1-\alpha_j\right) + \sum_{j=k+1}^{n} \lambda_j\right), \\
&\geq \min_U \left(\lambda_k \left[\sum_{j=1}^{k} \alpha_j - \sum_{j=k+1}^{n} 1 + \sum_{j=k+1}^{n} \alpha_j\right] + \sum_{j=k+1}^{n} \lambda_j\right), \\
&\geq \min_U \left(\lambda_k \left[\sum_{j=1}^{n} \alpha_j - (n-k)\right] + \sum_{j=k+1}^{n} \lambda_j\right), \\
&\geq \min_U \left(\lambda_k \left[\sum_{j=1}^{k} \alpha_j\right] - \lambda_k \left(n-k\right) + \sum_{j=k+1}^{n} \lambda_j\right), \\
&\geq \min_U \left(\lambda_k \left(n-k\right) - \lambda_k \left(n-k\right) + \sum_{j=k+1}^{n} \lambda_j\right), \\
&\geq \min_U \left(\sum_{j=k+1}^{n} \lambda_j\right).
\end{aligned}$$

If we set $M = \text{span}\{\varphi_1, \varphi_2, \ldots \varphi_k\}$), (i.e., $U = \Phi$) then it is easy to check that
Equation:

$$E[||x-x_M||^2] = \sum_{j=k+1}^{n} \lambda_j,$$

proving the theorem.

Example 3 (Transform Coding) Transform coding is a common scheme for data compression that leverages the Karhuenen-Loève transform. Examples include JPEG and MP3. In particular, JPEG can be broadly described as follows:

1. Take the image x and create tiles of size 8×8. We assume that the tiles are draws from a random variable X, i.e., the tiles $x_1, x_2 ... \in \mathbb{R}^{64}$ with $R_X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$
2. Compute the KLT of the tile random variable X from R_X by obtaining its eigendecomposition $R_X = \Phi \Lambda \Phi^T$.
3. Compute KLT coefficients for each block as $c_i = \Phi^T x_i$.
4. Pick as many coefficients of c_i as allowed by communications or storage constraints; save them as the compressed image.
5. Load saved coefficients and append zeros to build coefficient vector $\hat{c}_i$.
6. Run inverse KLT to obtain the decompressed tiles $\widehat{x}_i = \Phi \hat{c}_i$.
7. Reassemble the image from the decompressed tiles.

In practice, it is not desirable to recompute the KLT for each individual image. Thus, the JPEG algorithm employs the discrete cosine transform (DCT). It turns out that the DCT is a good approximation of the KLT for tiles of natural images. Additionally, instead of selecting a subset of the coefficients, they are quantized to varying quality/error according to their index and the total amount of bits available.

Singular Value Decomposition
Introduces the singular value decomposition and its application in principal component analysis.

Singular value decomposition (SVD) can be thought of as an extension to eigenvalue decomposition for non-symmetric matrices. Consider an $m \times n$ matrix X. The following two matrices are symmetric and so have eigenvalue decompositions

Equation:

$$XX^H = U\Lambda_1 U^H \text{ and } X^H X = V\Lambda_2 V^H,$$

where XX^H is an $m \times m$ matrix and $X^H X$ is an $n \times n$ matrix. It turns out that we can therefore decompose the matrix X as $X = U\Sigma V^H$, where Σ is an $m \times n$ “diagonal” matrix: $\Sigma_{i,i} = \sigma_i$ are the singular values of X, and $\Sigma_{i,j} = 0$ for $i \neq j$. The pseudoinverse of the matrix can then be written as $X^\dagger = V\Sigma^\dagger U^H$, where $\Sigma^\dagger_{i,i} = 1/\sigma_i$ and $\Sigma^\dagger_{i,j} = 0$ for $i \neq j$.

Principal Component Analysis

Principal component analysis can be thought of as KLT for sampled data. Assume that $\{x_1, x_2, \cdots x_L\} \subseteq \mathbb{R}^n$ is a zero-mean dataset, and collect it into a matrix $X = [x_1 x_2 \cdots x_L] \subseteq \mathbb{R}^{nxL}$. Next, compute the SVD $X = U\Sigma V^T$ with the corresponding eigenvalue decomposition $XX^T = U\Lambda U^T$. The matrix U is known as the principal component analysis (PCA) matrix of X; its columns $U_1, U_2, ...U_n$ are known as principal components, and its PCA coefficients are given by $Y = U^T X = \Sigma V^T$. The matrix Y contains the “scores” of all data points in the columns of X against the principal components U_i. One can show that the principal components in the matrix U follow the formulation

Equation:

$$\begin{aligned} u_j = \ & \underset{w}{\text{argmax}} \in \mathbb{R}^n \frac{1}{L} \sum_{i=1}^{L} |\langle x_i, u_i \rangle|^2 \\ & \text{subject to } \langle w, u_i \rangle = 0, \ i = 1, ..., j-1. \end{aligned}$$

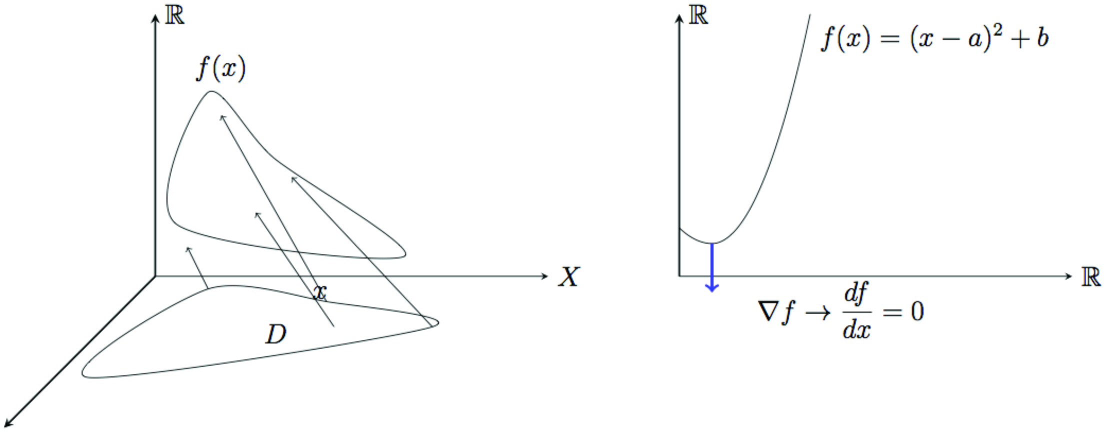
In words, u_j is the direction in which the projections of the data has the largest variance while being orthogonal to $\{u_1, u_2, ...u_{j-1}\}$.

Optimization in Hilbert Spaces
Introduces general optimization theory in Hilbert Spaces

In the remainder of the course we will discuss optimization problems. In general, an optimization problem consists of picking the “best” signal according to some metric; the metric will be some functional $f : X \to Y$ and the search will be over a set of interest $D \subseteq X$, so that the problem can be written as
Equation:

$$\widehat{x} = \underset{x\in D}{\operatorname{argmax}}\ f(x) \ \text{ or } \ \widehat{x} = \underset{x\in D}{\operatorname{argmin}}\ f(x).$$

We will need to extend the ideas of derivatives and gradients (which are used in optimization of single-variable real-valued functions) to arbitrary signal spaces where we can move in infinite directions on a set of interest.



Optimization examples. We wish to find the largest or smallest value of a functional $f(x)$, $x \in X$, over a set $D \subseteq X$. For scalar-valued functions of scalar fields, the maximizer/minimizer is found by solving $\frac{df}{dx} = 0$.

Directional Derivatives

Assume that we have a metric function $f : X \to Y$ and a set of interest $D \subseteq X$. Navigating the “surface” of f to find a maximum or minimum requires for us to formulate a framework for derivatives.

Definition 1 Let $x \in D \subseteq X$ and $h \in X$ be arbitrary. If the limit
Equation:

$$\delta f(x;h) = \lim_{\alpha\to 0} \frac{1}{\alpha}\left[f(x+\alpha h) - f(x)\right]$$

exists, it is called *Gâteaux differential* of f at x with increment (or in the direction) h. If the limit exists for each $h \in X$, the transformation f is said to be *Gâteaux differentiable* at x. If f is Gâteaux differentiable at all $x \in X$, then it is called a *Gâteaux differentiable functional*.

This extends the concept of derivative to incorporate direction so it can be used for any signal space. Note that α needs to be sufficiently small so that $x + \alpha h \in D$. Note also that for a fixed point x and variable direction h, the Gâteaux differential is a map from X to Y, i.e., $\delta f(x; \cdot) : X \to Y$.

Fact 1 In the common case of $Y = \mathbb{R}$,
Equation:

$$\delta f\left(x;h\right) = \frac{\partial}{\partial \alpha} f\left(x + \alpha h\right)\Big|_{\alpha=0}.$$

Example 1 Let H be a Hilbert space and $L \in B(H,H)$. Define the function $f : H \to \mathbb{R}$ by $f(x) = \langle Lx, x\rangle$. What is its Gâteaux differential? From the definition,
Equation:

$$\delta f\left(x;h\right) = \frac{\partial}{\partial \alpha}\left(\langle L\left(x + \alpha h\right), x + \alpha h\rangle\right)\Big|_{\alpha=0}.$$

We compute the derivative:
Equation:

$$\begin{aligned} \langle L(x+\alpha h), x+\alpha h\rangle &= \langle Lx, x\rangle + \alpha\,\langle Lh, x\rangle + \alpha\,\langle Lx, h\rangle + \alpha^2\,\langle Lh, h\rangle, \\ \frac{\partial}{\partial \alpha}\left(\langle L\left(x+\alpha h\right), x+\alpha h\rangle\right) &= \langle Lh, x\rangle + \langle Lx, h\rangle + 2\alpha\langle Lh, h\rangle. \end{aligned}$$

Therefore,
Equation:

$$\begin{aligned} \delta f(x;h) &= \langle Lh, x\rangle + \langle Lx, h\rangle = \langle Lh, x\rangle + \langle h, Lx\rangle = \left\langle h, L^* x\right\rangle + \langle h, Lx\rangle, \\ &= \left\langle h, \left(L + L^*\right) x\right\rangle. \end{aligned}$$

Unfortunately, the Gâbeaux differential does not satisfy our need to connect differentiability to continuity.

Definition 2 Let $f : X \to Y$ be a transformation on $D \subseteq X$. If for each $x \in D$ and each direction $h \in X$ there exists a function $\delta f(x; h) : D \times X \to Y$ that is linear and continuous with respect to h such that
Equation:

$$\lim_{\|h\| \to 0} \frac{\|\ f(x+h) - f(x) - \delta f(x;h)\ \|}{\|\ h\ \|} = 0,$$

then f is said to be *Fréchet differentiable* at x and $\delta f(x; h)$ is said to be the *Fréchet differential* of f at x with increment h.

One can intuitively see that there is a stronger connection between the common definition of a derivative (for functions $\mathbb{R} \to \mathbb{R}$) and the Fréchet derivative. There are additional connections between the derivatives and their properties.

Lemma 1 If a function f is Fréchet differentiable then $\delta f(x;h)$ is unique.

Lemma 2 If the Fréchet differential of f exists at x, then the Gâteaux differential of f exists at x and they are equal.

Lemma 3 If f defined on an open set $D \subseteq X$ has a Fréchet differential at x then f is continuous at x.

For a Fréchef-differentiable function, for any $\epsilon > 0$ there exists a sufficiently small $h \in X$ such that
Equation:

$$\frac{\parallel f(x+h) - f(x) - \delta f(x;h) \parallel}{\parallel h \parallel} < \epsilon.$$

This in turn implies
Equation:

$$\begin{aligned} \parallel f(x+h) - f(x) \parallel & \leq \parallel f(x+h) - f(x) - \delta f(x;h) \parallel + \parallel \delta f(x;h) \parallel \leq \epsilon \parallel h \parallel + \parallel \delta f(x;\cdot) \parallel \parallel h \parallel, \\ & \leq (\epsilon + \parallel \delta f(x;\cdot) \parallel) \parallel h \parallel, \end{aligned}$$

as $\delta f(x;h)$ is a linear continuous functional on h, implying that it is bounded. Therefore, as $\parallel h \parallel \rightarrow 0$, we have
Equation:

$$\lim_{\parallel h \parallel \rightarrow 0} \parallel f\left(x+h\right) - f\left(x\right) \parallel = 0.$$

This implies that f is continuous at x.

Local Optimization
Describes conditions for local optimization in Hilbert Spaces

We also must define the notion of an extremum in an arbitrary normed space.

Definition 1 Let f be a real-valued functional defined on $\Omega \subseteq X$ where X is a normed space. A point $x_0 \in \Omega$ is a *local/relative minimum* of f on Ω if $f\left(x_0\right) \leq f\left(x\right)$ for all $x \in \Omega$ such that $\parallel x - x_0 \parallel < \epsilon$ for some $\epsilon > 0$.

Definition 2 Let f be a real-valued functional defined on $\Omega \subseteq X$ where X is a normed space. A point $x_0 \in \Omega$ is a *local maximum* of f on Ω if $f\left(x_0\right) \geq f\left(x\right)$ for all $x \in \Omega$ such that $\parallel x - x_0 \parallel < \epsilon$ for some $\epsilon > 0$.

Definition 3 Let f be a real-valued functional defined on $\Omega \subseteq X$ where X is a normed space. A point $x_0 \in \Omega$ is a *local strict minimum* of f on Ω if $f\left(x_0\right) < f\left(x\right)$ for all $x \in \Omega$ such that $\parallel x - x_0 \parallel < \epsilon$ for some $\epsilon > 0$.

Definition 4 Let f be a real-valued functional defined on $\Omega \subseteq X$ where X is a normed space. A point $x_0 \in \Omega$ is a *local strict maximum* of f on Ω if $f\left(x_0\right) > f\left(x\right)$ for all $x \in \Omega$ such that $\parallel x - x_0 \parallel < \epsilon$ for some $\epsilon > 0$.

It turns out the notion of a gradient is intrinsically linked to the directional derivatives we have introduced.

Definition 5 Let X be a Hilbert space and $f : X \rightarrow R$. If f is a Fréchet differentiable functional, then for each $x \in X$ there exists a vector in X such that $\delta f(x; h) = \langle h, \nabla f(x) \rangle$ for all $h \in X$; the vector $\nabla f(x)$ is called the *gradient* of f at x, and can be written as a functional $\nabla f : X \rightarrow X$.

This definition can be seen to correspond to an application of the Riesz representation theorem to the Fréchet derivative $\delta f(x; h)$, which is a linear bounded functional on h.

Example 1 We know now that:
Equation:

$$\delta f(x; h) = \langle h, \nabla f(x) \rangle .$$

By the Cauchy-Schwarz Inequality, we have:
Equation:

$$|\delta f(x;h)| = |\langle h, \nabla f(x)\rangle| \leq \| h \| \| \nabla f(x) \|.$$

If $h = \nabla f(x)$ then $\delta f(x;h)$ is maximized.

Example 2 Recall that if $f : \mathbb{R}^n \to \mathbb{R}$, then
Equation:

$$\nabla f\left(x\right) = \begin{matrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{matrix} .$$

Theorem 1 Let $f : X \to \mathbb{R}$ have a Gâteaux differential on X. A necessary condition for f to have an extremum at $x_0 \in X$ is that $\delta f(x_0;h) = 0$ for all $h \in X$. Alternatively, if X is a Hilbert space, we can write $\langle h, \nabla f\left(x_0\right)\rangle = 0$ for all $h \in X$, which implies $\nabla f(x_0) = 0$.

Suppose x_0 is a local minimum. Then there exists $\epsilon > 0$ such that if $\| x - x_0 \| < \epsilon$ then $f\left(x_0\right) \leq f\left(x\right)$. Fix $h \neq 0$ and let $\theta = \frac{\epsilon}{\|h\|}$. Next, consider $x = x_0 + \alpha h$. For $\alpha \in (-\theta, \theta)$:

- If $\alpha > 0$ then $\frac{f(x_0+\alpha h) - f(x_0)}{\alpha} \geq 0$, and therefore $\delta f(x_0;h) \geq 0$.
- If $\alpha < 0$ then $\frac{f(x_0+\alpha h) - f(x_0)}{\alpha} \leq 0$, and therefore $\delta f(x_0;h) \leq 0$.

Therefore, $\delta f(x_0;h) = 0$ for arbitrary nonzero h. Now since $\delta f(x;h)$ is linear on h we must have $\delta f(x;h) = 0$ for $h = 0$. Therefore, the equality is true for all $h \in X$.

Definition 6 A point at which $\delta f(x; h) = 0$ for all $h \in X$ is called a *stationary point* of f.

Calculus of Variations
Introduces calculus of variations problems in optimization and their solutions based on the Euler-Lagrange equation.

The calculus of variations refers to a generic class of optimization problems that can be written in terms of
Equation:

$$\begin{array}{rl} \text{minimize} & J(x) \\ \text{subject to} & x\left(t_1\right)=a_1,\ x\left(t_2\right)=a_2, \end{array}$$

where $J\left(x\right):C\left[t_1,t_2\right]\to R$ is a function that can be written as
Equation:

$$J\left(x\right)=\int_{t_1}^{t_2} f\left(x\left(t\right),\dot{x}\left(t\right),t\right)dt,$$

where $\dot{x}\left(t\right)=\frac{dx(t)}{dt}$. The function f must meet the following conditions:

- $f(x\left(t\right),\dot{x}\left(t\right),t)$ is continuous on $x(t)$, $\dot{x}\left(t\right)$, and t as individual inputs,
- $f(x\left(t\right),\dot{x}\left(t\right),t)$ has continuous partial derivatives with respect to $x(t)$ and $\dot{x}\left(t\right)$, written as
  Equation:

$$\begin{aligned} f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right) &= \frac{\partial}{\partial x} f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right), \\ f_{\dot{x}}\left(x\left(t\right),\dot{x}\left(t\right),t\right) &= \frac{\partial}{\partial \dot{x}} f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right). \end{aligned}$$

Consider the set of admissible functions $x\in C[t_1,t_2]$. One can pick any particular admissible function x and then define the rest of the set in terms of $x+h$, where $h\left(t_1\right)=h\left(t_2\right)=0$. Therefore, at an optimum x, we require the directional derivative in the feasible directions h to be zero valued, i.e., $\partial J(x;h)=0$ for all feasible directions h. Now recall that since $J(x)$ is scalar-valued, we have
Equation:

$$\begin{aligned} \partial J(x;h) &= \frac{\partial}{\partial \alpha} J\left(x+\alpha h\right)|_{\alpha=0}, \\ &= \frac{\partial}{\partial \alpha}\int_{t_1}^{t_2}\left[f\Big(x\left(t\right)+\alpha h\left(t\right),\dot{x}\left(t\right)+\alpha\dot{h}\left(t\right),t\Big)\right]dt|_{\alpha=0}, \\ &= \int_{t_1}^{t_2}\frac{\partial}{\partial \alpha}\left[f\Big(x\left(t\right)+\alpha h\left(t\right),\dot{x}\left(t\right)+\alpha\dot{h}\left(t\right),t\Big)\right]dt|_{\alpha=0}. \end{aligned}$$

We use the following fact:

Fact 1 For any function $g(x,y,z)$, we have that
Equation:

$$\frac{\partial}{\partial \alpha} g\left(x+\alpha x_1,y+\alpha y_1,z\right)=\frac{\partial}{\partial x} g\left(x+\alpha x_1,y+\alpha y_1,z\right)x_1+\frac{\partial}{\partial y} g\left(x+\alpha x_1,y+\alpha y_1,z\right)y_1.$$

Therefore,
Equation:

$$\begin{aligned}\partial J(x;h) &= \int_{t_1}^{t_2}\left[\frac{\partial}{\partial x} f\left(x\left(t\right)+\alpha h\left(t\right), \dot{x}\left(t\right)+\alpha\dot{h}\left(t\right), t\right)h + \frac{\partial}{\partial \dot{x}} f\left(x\left(t\right)+\alpha h\left(t\right),\dot{x}\left(t\right)+\alpha\dot{h}\left(t\right),t\right)\dot{h}\right]dt|_{\alpha=}\\ &= \left[\int_{t_1}^{t_2} f_x\left(x+\alpha x,\dot{x}+\alpha\dot{h},t\right)hdt + \int_{t_1}^{t_2} f_{\dot{x}}\left(x+\alpha x,\dot{x}+\alpha\dot{h},t\right)\dot{h}dt\right]_{\alpha=0},\end{aligned}$$

where we drop the dependence of the functions for brevity (i.e., $x(t)$ is written x). Using integration by parts ($u = f_{\dot{x}}\left(x+\alpha x,\dot{x}+\alpha\dot{h},t\right), dv = \dot{h}dt$), we get

Equation:

$$\begin{aligned}\partial J(x;h) = \; &\left[\int_{t_1}^{t_2} f_x\left(x+\alpha x,\dot{x}+\alpha\dot{h},t\right)hdt + \left[f_x\left(x\left(t\right)+\alpha h\left(t\right),\dot{x}\left(t\right)+\alpha\dot{h}\left(t\right),t\right)h\left(t\right)\right]\Big|_{t=t_1}^{t_2}\right.\\ &\left. - \int_{t_1}^{t_2} h\frac{\partial}{\partial t} f_{\dot{x}}\left(x+\alpha h, x+\alpha\dot{h},t\right)dt\right]_{\alpha=0};\end{aligned}$$

since $h\left(t_1\right) = h\left(t_2\right) = 0$, we have that

Equation:

$$\begin{aligned}\partial J(x;h) &= \left\{\int_{t_1}^{t_2} h\left(t\right)\left[f_x\left(x+\alpha h,\dot{x}+\alpha h,t\right) - \frac{\partial}{\partial t} f_{\dot{x}}\left(x+\alpha h,\dot{x}+\alpha\dot{h},t\right)\right]dt\right\}_{\alpha=0},\\ &= \int_{t_1}^{t_2} h\left(t\right)\left[f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right) - \frac{\partial}{\partial t} f_{\dot{x}}\left(x\left(t\right),\dot{x}\left(t\right),t\right)\right]dt.\end{aligned}$$

It follows that for $\partial J(x;h) = 0$ for all feasible directions α, we must have

Equation:

$$f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right) - \frac{\partial}{\partial t} f_{\dot{x}}\left(x\left(t\right),\dot{x}\left(t\right),t\right) = 0,$$

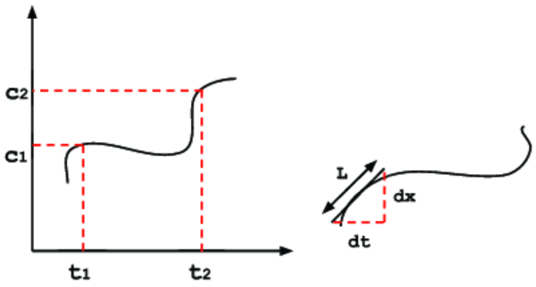
which provides the following condition on the solution $x(t)$ of the problem [link]:

Equation:

$$f_x\left(x\left(t\right),\dot{x}\left(t\right),t\right) = \frac{\partial}{\partial t} f_{\dot{x}}\left(x\left(t\right),x\left(t\right),t\right).$$

This condition is known as the *Euler-Lagrange* equation.

Example 1 Looking for a function $x\left(t\right) \in C\left[t_1,t_2\right]$ that minimizes the length of the path (curve) between the points $\left(t_1,c_1\right)$ and $\left(t_2,c_2\right)$, as shown in [link].



Example of a calculus of variations problem: finding the curve connecting two points that achieves minimum length, which is computed in a differential fashion.

In this case, we can consider increments of the curve's length l in terms of differences of the input dt and the output dx:
Equation:

$$dL = \sqrt{dt^2 + dx^2} = dt\sqrt{1 + \left(\frac{dx}{dt}\right)^2} dt\sqrt{1 + \dot{x}^2},$$

and by integrating both sides we get that
Equation:

$$L = \int_{t_1}^{t_2} dL = \int_{t_1}^{t_2} \sqrt{1 + \dot{x}^2} dt = \int_{t_1}^{t_2} f\left(x, \dot{x}, t\right) dt,$$

where we have written $f\left(x, \dot{x}, t\right) = \sqrt{1 + \dot{x}^2}$, and the problem has set the boundary conditions $x\left(t_1\right) = c_1$, $x\left(t_2\right) = c_2$. Thus, we have a calculus of variations problem.

To obtain the solution to this problem, we set up the Euler-Lagrange equation:
Equation:

$$\begin{aligned} f_x\left(x, \dot{x}, t\right) &= 0, \\ f_{\dot{x}}\left(x, \dot{x}, t\right) &= \frac{1}{2} \cdot \left(1 + \dot{x}\right)^{-\frac{1}{2}} \cdot 2\dot{x} = \frac{\dot{x}}{\sqrt{1 + \left(\dot{x}\right)^2}}, \\ 0 &= \frac{d}{dt} f_{\dot{x}}\left(x, \dot{x}, t\right); \end{aligned}$$

in words, $f_{\dot{x}}\left(x, \dot{x}, t\right)$ must be a constant as a function of t, which implies that $\dot{x}\left(t\right)$ must be a constant function of t; such a function with constant first derivative is a straight line. Therefore, the shortest path between the two aforementioned points is obtained by the straight line that connects them.

Example 2 Consider a retirement plan with the following constraints:

- Your current capital is S dollars, and ideally by the end of your life you will have spent it all; that is, if $x(t)$ is your capital at time t, then $x(0) = S$ and $x(T) = 0$.
- Your expense rate is given by the function $r(t)$, and spending r dollars gives you a quantifiable amount of enjoyment $u[r(t)]$.

The goal of the planning is to maximize your total enjoyment:
Equation:

$$\int_0^T e^{-\beta t} u\left[r\left(t\right)\right] dt,$$

where the exponential weights enjoyment to specify that enjoyment decreases with age. The change in your capital is given by its derivative, which must account for your expense rate and the return on investment:
Equation:

$$\dot{x} = -r\left(t\right) + \alpha x\left(t\right),$$

where $\alpha > 1$. The problem is thus to maximize the function of your capital function
Equation:

$$J\left(x\right) = \int_0^T e^{-\beta t} u\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right] dt = \int_0^T f\left(x\left(t\right), \dot{x}\left(t\right), t\right) dt,$$

which together with the initial constraints gives us a calculus of variation problem. Thus, once again, we set up the Euler-Lagrange equation: if we denote $u'\left[r\right] = \frac{d}{dr} u\left[r\right]$, then
Equation:

$$\begin{aligned} f_x\left(x, \dot{x}, t\right) &= \alpha e^{-\beta t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right], \\ f_{\dot{x}}\left(x, \dot{x}, t\right) &= e^{-\beta t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right], \\ \frac{\partial}{\partial t} f_{\dot{x}}\left(x, \dot{x}, t\right) &= \beta e^{-\beta t} u'\left[\alpha x\left(t\right) - \dot{x}\right] - e^{-\beta t} \frac{\partial}{\partial t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right]. \end{aligned}$$

So we obtain
Equation:

$$\begin{aligned} \alpha e^{-\beta t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right] &= \beta e^{-\beta t} u'\left[\alpha x\left(t\right) - \dot{x}\right] - e^{-\beta t} \frac{\partial}{\partial t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right], \\ (\beta - \alpha) u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right] &= \frac{\partial}{\partial t} u'\left[\alpha x\left(t\right) - \dot{x}\left(t\right)\right]. \end{aligned}$$

Now we can switch back to the rate of expense $r\left(t\right) = \alpha x\left(t\right) - \dot{x}\left(t\right)$ to get
Equation:

$$(\beta - \alpha) u'\left[r\left(t\right)\right] = \frac{d}{dt} u'\left[r\left(t\right)\right],$$

which is a differential equation. The solution for $u'\left[r\left(t\right)\right]$ is therefore given by
Equation:

$$u'\left[r\left(t\right)\right] = u'\left[r\left(0\right)\right] e^{(\beta - \alpha) t}.$$

To move forward, we need to select a candidate form for the utility function u. Our goals for this function is to showcase a diminishing marginal enjoyment as one spends increasing amounts of money (i.e., $u'\left[m\right]\to 0$ as $m\to\infty$) and a sense of significantly increasing enjoyment as the amount of money spent is small but increasing (i.e., $u'\left(0\right)=\infty$). A candidate function that obeys these two conditions is $u\left[m\right]=2m^{1/2}$, which provides $u'\left[m\right]=m^{-1/2}$; replacing in [link], we get that the rate of expense must obey

Equation:

$$\begin{aligned} r(t)^{-1/2} &= r(0)^{-1/2}e^{(\beta-\alpha)t}, \\ r(t) &= r\left(0\right)e^{2(\alpha-\beta)t}. \end{aligned}$$

Connecting back to the capital function, we get that

Equation:

$$\alpha x\left(t\right)-\dot{x}\left(t\right)=r\left(0\right)e^{2(\alpha-\beta)t},$$

which is another differential equation. The solution to this equation is

Equation:

$$x\left(t\right)=e^{\alpha t}x\left(0\right)+\frac{r(0)}{\alpha-2\beta}\left(e^{\alpha t}-e^{2(\alpha-\beta)t}\right).$$

In this equation we can replace $x(0)=S$; assuming $\alpha>\beta>\alpha/2$, we can find $r(0)$ by setting $T=0$ above; since $x(T)=0$, we get

Equation:

$$r\left(0\right)=\frac{(2\beta-\alpha)x(0)}{1-e^{(\alpha-2\beta)T}}=\frac{(2\beta-\alpha)S}{1-e^{(\alpha-2\beta)T}}.$$

Therefore, the final solution to the problem is

Equation:

$$\begin{aligned} x(t) &= e^{\alpha t}S-\frac{S}{1-e^{(\alpha-2\beta)T}}\left(e^{\alpha t}-e^{2(\alpha-\beta)t}\right), \\ &= e^{\alpha t}S\left(\frac{e^{(\alpha-2\beta)t}-e^{(\alpha-2\beta)T}}{1-e^{(\alpha-2\beta)T}}\right). \end{aligned}$$

Constrained Optimization

Introduces the theory of constrained optimization, including Lagrange multipliers.

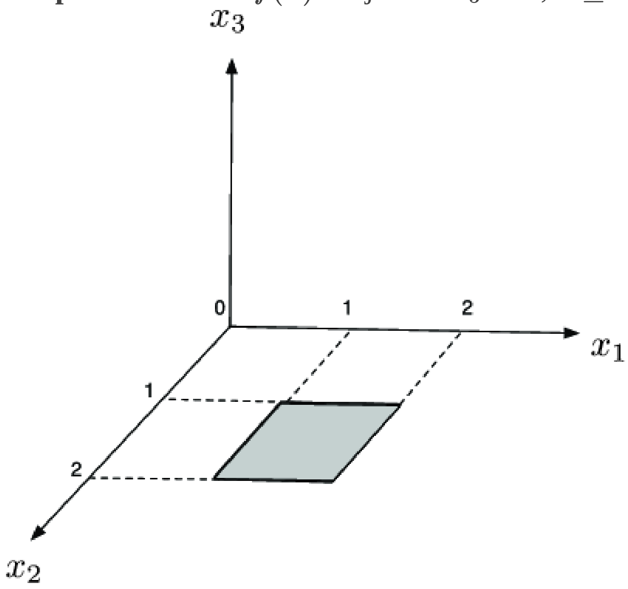
In constrained optimization, we look to minimize or maximize an objective function only over a set of inputs $\Omega \subseteq X$ that can be written in the following form:

Equation:

$$\Omega = \left\{ \begin{array}{lll} x_0 \in X: & g_1(x) = 0, & h_1(x) \leq 0, \\ & g_2(x) = 0, & h_2(x) \leq 0, \\ & g_3(x) = 0, & h_3(x) \leq 0, \\ & \vdots & \vdots \end{array} \right\}$$

In words, Ω is a set described by set of equalities and inequalities on x. The constraints $g_i(x) = 0$ are said to be *equality constraints*, and the constraints $h_i(x) \leq 0$ are said to be *inequality constraints*.

Example 1 Minimize $f(x)$ subject to $x_3 = 0$, $1 \leq x_1 \leq 2$ and $1 \leq x_2 \leq 2$, drawn on [link].



An example of a feasible set Ω that can be expressed using equality and inequality constraints.

In this optimization problem, the feasible set Ω can be written in terms of one equality constraint, $g(x) = x_3$, and four inequality constraints:

Equation:

$$\begin{aligned} h_1(x) &= 1 - x_1, \\ h_2(x) &= x_1 - 2, \\ h_3(x) &= 1 - x_2, \\ h_4(x) &= x_2 - 2. \end{aligned}$$

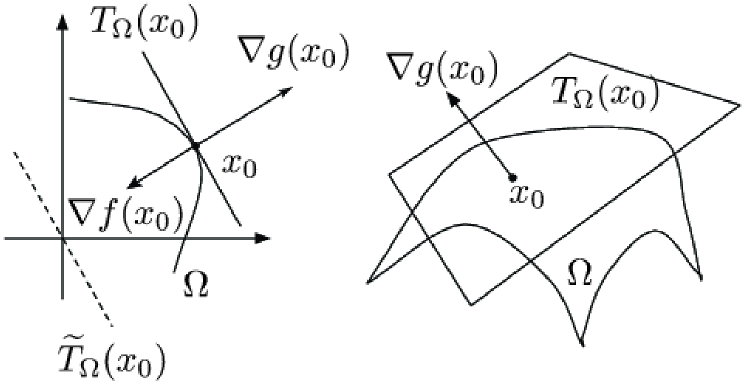
Definition 1 The points $x \in \Omega$ are called *feasible points*, and the set Ω is called the *feasible set*.

In the sequel, we will assume that f, g_i, h_i are continuous and Fréchet-differentiable (continuous gradients).

We will first consider problems where the feasible set Ω can be expressed in terms of equality constraints only.

Tangent space

For a given feasible point $x_0 \in \Omega$, the *tangent space* gives us the set of directions in which one can move from x_0 while still staying within the feasible set Ω. The two examples below show tangent spaces in the cases where the set Ω correspond to a curve in $\mathbb{R}^2$ and a nonlinear manifold in $\mathbb{R}^3$.



Examples of tangent spaces and gradients.

The tangent space at x_0 can be expressed formally in terms of the derivatives of the equality constraints.

Definition 2 The *tangent space* to the feasible set Ω with equality constraints $g_1, ..., g_n$ at a feasible point $x_0 \in \Omega$ is given by
Equation:

$$\begin{aligned} T_\Omega(x_0) &= \{d \in X : \delta g_i(x_0; d - x_0) = 0,\ i = 1, 2, \cdots, n\}, \\ &= \{d \in X : \langle \nabla g_i(x_0), d - x_0 \rangle = 0,\ i = 1, 2, \cdots, n\}. \end{aligned}$$

Requiring all the directional derivatives of the equality constraints to be zero in the direction of the tangent $d - x_0$ guarantees that the value of the equality constraints remains at zero, therefore guaranteeing that d remains a feasible point, i.e., $d \in \mathit{\Omega}$.

Definition 3 A point x_0 satisfying the constraints $g_1\left(x_0\right) = 0, g_2\left(x_0\right) = 0, ..., g_n\left(x_0\right) = 0$ is said to be a regular point if the n linear functionals $\delta g_1\left(x_0; h\right), \delta g_2\left(x_0; h\right), ..., \delta g_n\left(x_0; h\right)$ (i.e., the derivatives of the equality constraints) are linearly independent on h.

Theorem 1 If x_0 is an extremum of $f(x)$ subject to constraints $g_1\left(x_0\right) = 0, g_2\left(x_0\right) = 0, ..., g_n\left(x_0\right) = 0$ and x_0 is a regular point of $\left\{g_i\right\}_{i=1}^{n}$ then for any $h \in X$ such that $\delta g_i\left(x; h\right) = 0$ for all $i = 1, 2, ..., n$, we must have $\delta f(x_0; h) = 0$.

One can rewrite this theorem in terms of gradients as: if $\langle \nabla g_i\left(x_0\right), h\rangle = 0$, then $\langle \nabla f\left(x_0\right), h\rangle = 0$. More intuition can be obtained by defining the translated tangent space at x_0 as follows:

Equation:

$$\widetilde{T}_{\mathit{\Omega}}\left(x_0\right) = \left\{d \in X : \langle \nabla g_i\left(x_0\right), d\rangle = 0, i = 1, 2, \cdots, n\right\}.$$

Thus, the theorem above can be written as: if $h \in \widetilde{T}_{\mathit{\Omega}}\left(x_0\right)$, then $\langle \nabla f\left(x_0\right), h\rangle = 0$. In words, we expect for the derivative of the objective function f at the constrained extremum x_0 to be zero-valued in the directions in which we can move from x_0 and remain in the feasible set $\mathit{\Omega}$ — that is, in the directions $h \in \widetilde{T}_{\mathit{\Omega}}\left(x_0\right)$. Thus, we can write that the constrained optimum x_0 must obey $\nabla f\left(x_0\right) \perp \widetilde{T}\left(x_0\right)$, which implies $\nabla f\left(x_0\right) \perp T_{\mathit{\Omega}}\left(x_0\right)$.

Example 2 Let $X = \mathbb{R}^2$; solve

Equation:

$$x_0 = \arg \max_{x = \left[x_1\ x_2\right]^T} f\left(x\right) = x_1 + x_2 \ \text{subjectto}\ x_1^2 + x_2^2 = 1.$$

In words, we look for the point in a unit circle in $\mathbb{R}^2$ that has the largest sum of its coordinates. It is easy to see by inspection that such point is given by $x_0 = \left[\sqrt{2}\ \sqrt{2}\right]^T$. The feasible set can be given in terms of the single equality constrain

Equation:

$$g\left(x\right) = x_1^2 + x_2^2 - 1 = x^T x - 1 = \langle x, x\rangle - 1 = \langle x, Ix\rangle - 1.$$

From previous work, we know that the gradient is given by $\nabla g\left(x\right) = \left(I + I^*\right)x = 2Ix = 2x$ We can also write the objective function as

Equation:

$$f(x) = \left\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, x \right\rangle,$$

which has gradient $\nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Thus, we can write the tangent space in this case as

Equation:

$$T_{\Omega}(x_0) = \{h \in X : \langle \nabla g(x_0), h \rangle = 0\} = \{h \in X : \langle x_0, h \rangle = 0\} = \left\{h : \sqrt{2}h_1 + \sqrt{2}h_2 = 0\right\}.$$

Therefore, we can write the tangent space as $T_{\Omega}(x_0) = \{h \in X : h_1 = -h_2\}$. It is easy to see at this point that for any such $h \in T_{\Omega}(x_0)$ we will have $\langle \nabla f(x_0), h \rangle = 0$, as stated by Theorem [link].

Lagrangian Multipliers

In this section, we will develop a method to solve optimization problems with linear objective functions and linear equality constraints.

Lemma 1 Let $f_0, f_1, f_2, ..., f_n$ be linear functionals on a Hilbert space X and suppose $f_0(x) = 0$ for all $x \in X$ such that $f_i(x) = 0,\ i = 1, 2, ..., n$. Then there exists constants $\lambda_1, \lambda_2, ..., \lambda_n$ such that $f_0 = \lambda_1 f_1 + \lambda_2 f_2 + ... + \lambda_n f_n$.

Since our functionals are linear we have $f_0, f_1, ..., f_n \in X^*$. Define the subspace $M = \text{span}(\{f_1, f_2, ..., f_n\})$. Since M is finite-dimensional, then M is closed. We can therefore define its orthogonal complement:

Equation:

$$M^{\perp} = \left\{f \in X^* : \langle f, f_i \rangle = 0,\ i = 1, 2, ..., n\right\}.$$

Since Hilbert spaces are self-dual, then for each function f_i there exists $w_i \in X$ such that $f_i(x) = \langle x, w_i \rangle$; therefore, we can rewrite the space above as its dual equivalent

Equation:

$$M^{\perp} = \{w \in X : \langle w, w_i \rangle = 0,\ i = 1, 2, ..., n\}.$$

Now since $\langle w, w_i \rangle = f_i(w)$, it follows that for all $w \in M^{\perp}$ we have that $f_i(w) = 0$, $w = 1, ..., n$. Therefore, the Lemma implies that for all $w \in M^{\perp}$ we have that $f_0(w) = 0 = \langle w, w_0 \rangle$. This implies that $w_0 \in \left(M^{\perp}\right)^{\perp} = M$, due to M being closed. Reversing to the dual space X^*, this implies that $f_0 \in M = \text{span}(\{f_1, ..., f_n\})$, and so we can write $f = \lambda_1 f_1 + \lambda_2 f_2 + ... + \lambda_n f_n$.

Theorem [link] shows that we can apply Lemma [link] to the constrained optimization problem. Thus, at the extremum $x_0 \in X$ of the constrained program there exist constants $\lambda_1, ..., \lambda_n$ such

that for all $h \in X$,
Equation:

$$\begin{aligned} \delta f(x_0;h) &= \sum_{i=1}^{n} \lambda_i \delta g_i\left(x_0;h\right), \\ \langle \nabla f\left(x_0\right), h\rangle &= \sum_{i=1}^{n} \lambda_i \left\langle \nabla g_i\left(x_0\right), h\right\rangle, \\ \left\langle \nabla f\left(x_0\right) - \sum_{i=1}^{n} \lambda_i \nabla g_i\left(x_0\right), h\right\rangle &= 0, \end{aligned}$$

which is equivalent to $\nabla f\left(x_0\right) + \sum_{i=1}^{n} g_i\left(x_0\right) = 0$. This can be written as the gradient of the Lagrangian function
Equation:

$$L\left(x,\lambda\right) = f\left(x\right) + \sum_{i=1}^{n} \lambda_i g_i\left(x\right).$$

Thus, we say that the extremum must provide a zero-valued directional derivative of the Lagrangian in all directions or, equivalently, a zero-valued gradient for the Lagrangian. The results obtained in this section can be collected into a proof for the following theorem.

Theorem 2 If $x_0 \in X$ is an extremum of a functional f subject to constraints $\{g_i\}_{i=1}^{n}$, then there exist scalars $\lambda_1, \cdots, \lambda_n$, such that the Lagrangian $L\left(x,\lambda\right) = f\left(x\right) + \sum_{i=1}^{n} \lambda_i g_i\left(x\right)$ is stationary at x_0, i.e., for all $h \in X$ we have that $\delta L(x,\lambda;h) = 0$, i.e., $\nabla L(x,\lambda) = 0$.

The constants $\lambda_1, \lambda_2, \cdots, \lambda_n$ are known as Lagrange multipliers.

Example 3 We want to minimize $f\left(x\right) = x_1^2 + x_2^2$ subject to $2x_1 + x_2 = 3$. The constraint function is
Equation:

$$g\left(x\right) = 2x_1 + x_2 - 3.$$

From earlier we know that $\nabla f(x) = 2x$, while we can rewrite the constraint as
Equation:

$$g\left(x\right) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} x + 3,$$

so that $\nabla g\left(x\right) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. Therefore, the extremum's condition on the gradient of the Lagrangian results in the equation

Equation:

$$\begin{aligned}\nabla f(x) + \lambda \nabla g(x) &= 0,\\ 2x + \lambda \begin{bmatrix} 2 \\ 1 \end{bmatrix} &= \mathbf{0},\\ \begin{bmatrix} 2x_1 + 2\lambda \\ 2x_2 + \lambda \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix};\end{aligned}$$

the solution to this equation is $x_1 = -\lambda, x_2 = -\frac{1}{2}\lambda$. To solve for the value of the Lagrangian multiplier λ, we plug the solution into the constraint: Plug in the constraint function
Equation:

$$2\left(-\lambda\right) + \left(-\frac{1}{2}\lambda\right) - 3 = 0,$$

which gives $\lambda = -\frac{6}{5}$. Therefore, we end up with the solution $x_1 = \frac{6}{5}, x_2 = \frac{3}{5}$.

Second Order Conditions
Describes second order conditions in local optimization to find maxima and minima.

When the objective function and constraints $f: \ \mathbb{R}^n \to \mathbb{R}$, $g_i : \ \mathbb{R}^n \to \mathbb{R}$, it is easy to check whether an extremum x_0 is a maximum or a minimum of the functional. We appeal to second-order differentials, known as Hessians.

Definition 1 The $n \times n$*Hessian matrix*$F(x)$ for the functional $f\left(x\right) : \mathbb{R}^n \to \mathbb{R}$ has entries
Equation:

$$F(x)_{ij} = \frac{\partial^2 f\left(x\right)}{\partial x_i \partial x_j}, \qquad i, j = 1, 2, \cdots, n.$$

Lemma 1 Let $\mathscr{L}(x)$ be the Hessian of the Lagrangian $L(x, \lambda)$ and let x_0 be an extremum. If $d^T \mathscr{L}\left(x_0\right) d \geq 0$ for all $d \in \widetilde{T}_\Omega\left(x_0\right)$, then x_0 is a minimizer. If $d^T \mathscr{L}\left(x_0\right) d \leq 0$ for all $d \in \widetilde{T}_\Omega\left(x_0\right)$, then x_0 is a maximizer.

Example 1 Find the extremum of $f\left(x\right) = {x_2}^2 + x_2 x_3 + x_1 x_3$ subject to $x_1 + x_2 + x_3 = 3$ and determine whether it is a maximum or a minimum.

To begin, we write the optimization's equality constraint:
Equation:

$$g\left(x\right) = x_1 + x_2 + x_3 - 3$$

The objective function can be written in the form
Equation:

$$f\left(x\right) = \sum_{ij} a_{ij} x_i x_j = x^T A x,$$

where the matrix A has entries given by $A_{ij} = a_{ij}$. Thus, for our example the resulting matrix is

Equation:

$$A = \begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{matrix} .$$

The gradient for this function is given by
Equation:

$$\nabla f\left(x\right) = \left(A + A^{*}\right)x = \begin{matrix} 0 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{matrix} x = Bx,$$

where $B = A + A^{*}$. We can also rewrite the inequality constraint as $g\left(x\right) = \mathbf{1}^{T}x - 3$, where $\mathbf{1}$ denotes a vector with entries equal to one of appropriate size. Therefore, its gradient is equal to $\nabla g(x) = \mathbf{1}$. The resulting gradient of the Lagrangian is set to zero to obtain the solution:
Equation:

$$\nabla f(x) + \lambda \nabla g(x) = 0$$

Equation:

$$Bx + \lambda \mathbf{1} = 0$$

Equation:

$$Bx = -\lambda \mathbf{1}$$

Equation:

$$x = -\lambda B^{-1}\mathbf{1} = \begin{matrix} -\lambda \\ 0 \\ -\lambda \end{matrix}$$

Solve for λ from $\underline{1}^T x = 3$ to obtain $\lambda = -3/2$. Therefore, the optimization's solution is

Equation:

$$x_0 = \begin{matrix} 3/2 \\ 0 \\ 3/2 \end{matrix} .$$

We can solve for the Hessians of $F(x)$ and $G(x)$:

Equation:

$$\begin{aligned} F(x) &= \left(A + A^*\right) = \begin{matrix} 0 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{matrix} , \\ G(x) &= \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} . \end{aligned}$$

We therefore obtain that the Hessian of the Lagrangian is equal to

Equation:

$$\mathscr{L}\left(x_0\right) = F\left(x_0\right) + \lambda G\left(x_0\right) = \begin{matrix} 0 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{matrix} = B.$$

At this point we need to check if the product $h^T \mathscr{L}\left(x_0\right) h$ is positive or negative for all $h \in \widetilde{T}_\Omega\left(x_0\right)$, the tangent space defined as

Equation:

$$\widetilde{T}_\Omega\left(x_0\right) = \{h : \langle \nabla g(x), h \rangle = 0\} = \{h : \langle 1, h \rangle = 0\}.$$

It is easy to see that $h \in \widetilde{T}_{\Omega}\left(x_0\right)$ if and only if $h_1 + h_2 + h_3 = 0$. To begin, we check whether the eigenvalues of $\mathscr{L}\left(x_0\right)$ are all positive or negative: a calculation returns $\{-1.\,1701, 0.\,6889, 2.\,4812\}$. Since neither case occurred, we have to specifically consider the case in which $h_1 + h_2 + h_3 = 0$:

Equation:

$$\begin{aligned} h^T \mathscr{L}\left(x_0\right) h &\gtrless 0, \\ \sum_{i,j=1}^{3} B_{ij} h_i h_j &\gtrless 0, \\ 2{h_2}^2 + 2h_2 h_3 + 2h_1 h_3 &\gtrless 0, \\ h_2^2 + h_3\left(h_1 + h_2\right) &\gtrless 0, \\ h_2^2 - h_3^2 &\gtrless 0. \end{aligned}$$

It turns out that we can find $h \in \widetilde{T}_{\Omega}\left(x_0\right)$ for which the value on the left hand side may be positive or negative. Therefore, this is neither a maximum or a minimum, and we have found an inflection point.

Constrained Optimization with Inequality Constraints
Introduces constrained optimization with inequality constraints and the Karush Kuhn Tucker conditions.

A constrained optimization problem with inequality constraints can be written as

Equation:

$$\min f(x) \ \text{ subject to } \ \begin{array}{l} g_i\left(x\right)=0,\ i=\left(1,\cdots,n\right), \\ h_j\left(x\right)\le 0,\ j=\left(1,\cdots,m\right). \end{array}$$

Definition 1 Let x_0 be a feasible point. If $h_j\left(x_0\right)=0$, we say that the constraint h_j is *active* at x_0, if $h_j\left(x_o\right)<0$ then h_j is *inactive* at x_0. The set of active constraints at x is denoted by;

Equation:

$$J\left(x\right)=\left\{j:\ h_j\left(x\right)=0\right\}$$

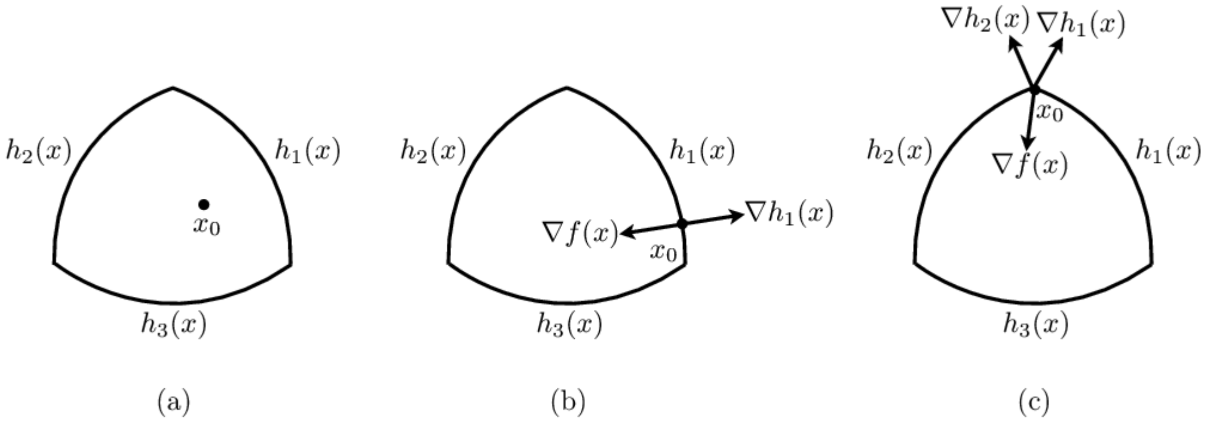
Definition 2 We say $x\in\varOmega$ is a *regular point* in $\varOmega$ if $\left\{\nabla g_i\left(x\right),\ i=1,\cdots,n\right\}\cup\left\{\nabla h_j\left(x\right),\ j\in J(x)\right\}$ is a linearly independent set.

Theorem 1 (Karush-Kuhn-Tucker Conditions) if x_0 is a regular point of $\varOmega$ and x_0 is a local minimizer of f over $\varOmega$ then there exist scalars $\lambda_1,\cdots,\lambda_n$ and $\mu_1,\cdots,\mu_m$ such that;

1. $\mu_j\ge 0$,$j=1,\cdots,m$,
2. $\sum_{j=1}^{m}\mu_j h_j\left(x_0\right)=0$ (complementary slackness condition),
3. $\nabla f(x_0)+\sum_{i=1}^{n}\lambda_i\nabla g_i\left(x_0\right)+\sum_{j=1}^{m}\mu_j\nabla h_j\left(x_0\right)=0.$

Since $h_j\left(x_0\right)\le 0$, we have that if h_j is inactive at x_0 (i.e., if $h_j\left(x_0\right)<0$) then we must have $\mu_j=0$. Therefore, some versions of the theorem feature only the active inequality constraints in the third condition.

Example 1 We pictorially demonstrate some examples of active inequality constraints: consider the case where the set $\varOmega$ is a convex set bounded by three inequality constraints, $h_1\left(x\right)$, $h_2\left(x\right)$, and $h_3\left(x\right)$. Now, consider these three possibilities for the minimizer x_0 of $f(x)$ in $\varOmega$, illustrated in [link]:



Three examples of the application of the Karhush-Kuhn-Tucker conditions.

- **(a)**In this case, the minimizer is in the interior of the set Ω, and no constraints are active; therefore, the minimizer of the constrained problem matches the minimizer of the unconstrained problem, and the solution is found by solving $\nabla f(x_0) = 0$, which ignores all inequality constraints. This means that each $\mu_i = 0$.
- **(b)**In this case, the minimizer has one active constraint ($h_1(x)$). Consider the gradient of the constraint $\nabla h_1(x)$, which is orthogonal to the tangent space. If $\nabla h_1(x)$ and $\nabla f(x)$ are not collinear (scalar multiples of one another), then there is a direction within the tangent space in which the value of $f(x)$ would decrease and x_0 would not be a minimizer. Otherwise, if $\nabla f(x) + \mu_1 \nabla h_1(x) = 0$ for $\mu_1 \leq 0$, then both gradients point in the same direction and the value of $f(x)$ would decrease by moving in the opposite direction toward the interior of Ω, and so x_0 would not be a minimizer. Therefore, we must have $\nabla f(x_0) + \mu_1 \nabla h_1(x_0) = 0$ with $\mu_1 > 0$, as specified by the KKT conditions. In this case the inactive constraints $h_2(x)$ and $h_3(x)$ are ignored.

- **(c)**In this case, the minimizer has two active constraints ($h_1\left(x\right)$ and $h_2\left(x\right)$). In this case, the directions d in which we can move from x_0 within ω must obey $\langle d,\nabla h_1\left(x_0\right)\rangle<0$ and $\langle d,\nabla h_2\left(x_0\right)\rangle<0$. Similarly, moving in a direction d decreases the value of $f(x_0+d)$ below $f(x_0)$ if $\langle d,\nabla f\left(x_0\right)\rangle<0$. Thus, for us to be able to find such a direction d we must have that $\nabla f(x_0)=ah_1\left(x_0\right)+b\nabla h_2\left(x_0\right)$ with $a\geq0$ and $b\geq0$, which would give
  Equation:
  $$\nabla f(x_0)-ah_1\left(x_0\right)-b\nabla h_2\left(x_0\right)=0.$$
  Thus, the minimizer x_0 of $f(x)$ must obey
  Equation:
  $$\nabla f(x_0)+\mu_1h_1\left(x_0\right)+\mu_2\nabla h_2\left(x_0\right)=0,$$
  with $\mu_1>0$ and $\mu_2>0$.

The example illustrates a simple “recipe” for solving inequality constraint optimization problems includes the following steps:

1. Pick a candidate active set $\widehat{J}\left(x_0\right)$,
2. Build the corresponding form of the third KKT condition:
   Equation:
   $$\nabla f(x_0)+\sum_{i=1}^{n}\lambda_i\nabla g_i\left(x_0\right)+\sum_{j\in\widehat{J}(x)}^{m}\mu_j\nabla h_j\left(x_0\right)=0,$$
   and solve for x_0,λ, and μ,
3. If $\mu_j\geq0$ and $h_j\left(x_0\right)=0$ for $j\in\widehat{J}\left(x_0\right)$ then x_0 is a solution. Otherwise, pick a new candidate active set $\widehat{J}\left(x\right)$ and repeat Step 2.

With some additional assumptions, it can be shown that the KKT conditions can find a global minimizer.

Definition 3 A function f is said to be *affine* over Ω if $f\left(\sum_i^n a_ix_i\right)=\sum_1^n a_if\left(x_i\right)$ for all $x_1,...,x_n\in\Omega$ and all weights $\{a_i\}$ obeying $\sum_i^n a_i=1$.

Theorem 2 (Karush-Kuhn-Tucker Sufficient Conditions) If f and h_j, $j=1,...,m$ are convex functions and g_i, $i=1,...,n$ are affine functions, and if the KKT condition are satisfied at a feasible point $x_0\in\Omega$ then x_0 is a global minimizer of f over Ω.

Fix $x_1 \in \mathit{\Omega}$ let $d = x_1 - x_0$. Define a functional $x\left(t\right) = tx_1 + (1-t)x_0 = x_0 + td$ over $t \in [0,1]$. Then, define the constraints limited over the set of points $x(t)$:
Equation:

$$\begin{aligned} G_i\left(t\right) &= g_i\left(x\left(t\right)\right) = g_i\left(tx_1 + (1-t)x_0\right) = t\; g\left(x_i\right) + (1-t)g\left(x_0\right) = 0, \\ H_j\left(t\right) &= h_j\left(x\left(t\right)\right) = h_j\left(tx_1 + (1-t)x_0\right) \leq th_j\left(x_1\right) + (1-t)h_j\left(x_0\right) \leq 0; \end{aligned}$$

Therefore, all points $x(t) \in \mathit{\Omega}$ are feasible. Furthermore, note that $H_j\left(0\right) = h_j\left(x_0\right) = 0 \geq H_j\left(t\right) = h_j\left(x_t\right)$ if $j \in J(x_0)$. Now, we compute the derivatives of these two functions with respect to t:
Equation:

$$\frac{\partial G}{\partial t} = 0 = \partial g_i\left(x_0, d\right) = \langle \nabla g_i\left(x_0\right), d\rangle,$$

and for $j \in J(x_0)$,
Equation:

$$0 \;\geq\; \frac{\partial H_j}{\partial t} = \partial h_j\left(x_0, d\right) = \langle \nabla h_j\left(x_0\right), d\rangle.$$

Now consider the function $F(t) = f(x(t))$: its derivative is given by
Equation:

$$\frac{\partial F}{\partial t} = \partial f\left(x_0, d\right) = \langle \nabla f\left(x_0\right), d\rangle = -\sum_{i=1}^{n} \langle \nabla g_i, d\rangle x_i - \sum_{i=1}^{n} \mu_j \left\langle \nabla h_j, d\right\rangle \geq 0,$$

where we use the third KKT condition. Since $f(x)$ is convex and $x(t)$ is affine, then $F(t) = f(x(t))$ is convex in $t \in [0,1]$. Thus $\frac{\partial F}{\partial t}$ is nondecreasing and $\frac{\partial F(t)}{\partial t} \geq \frac{\partial F(0)}{\partial t} \geq 0$ for $t \in [0,1]$. Thus, $F(1) \geq F(0)$ or $f\left(x_1\right) \geq f\left(x_0\right)$. Since x_1 was arbitrary, x_0 is a global minimum of f on $\mathit{\Omega}$.

Example 2 (Channel Capacity) The Shannon capacity of an additive white Gaussian noise channel is given by $C = \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right)$, where P is the transmitted signal power and N is the noise variance. Assume that n channels are available with a total transmission power $P_T = \sum_{i=1}^{n} P_i$ available among the channels, where P_i denotes the power in the i^{th} channel. We wish to assign a power profile $P = \left[P_1, ..., P_n\right]^T$ that maximizes the total capacity for the set of channels

Equation:

$$C\left(P\right)=\sum_{i=1}^{n}C\left(P_i\right)=\sum_{i=1}^{n}\frac{1}{2}\,\log_2\left(1+\frac{P_i}{N_i}\right),$$

where N_i represents the variance of the noise in the i^{th} channel.

To solve the problem, we set up an objective function to be minimized
Equation:

$$f\left(P\right)=-C\left(P\right)=-\sum_{i=1}^{n}C\left(P_i\right)=-\sum_{i=1}^{n}\frac{1}{2}\,\log_2\left(1+\frac{P_i}{N_i}\right)$$

and also set up the constraints
Equation:

$$\begin{aligned} g(P) &= \sum_{i=1}^{n}P_i-P_T=\mathbf{1}^TP-P_T,\\ h_i\left(P\right) &= -P_i=-e_i^TP,\ i=1,...,n, \end{aligned}$$

as the values of the powers must be nonnegative. We start by computing the gradients of these functions: for f, we must compute the directional derivative
Equation:

$$\begin{aligned} \delta f(p;h) &= \frac{\partial}{\partial\alpha}\left(f\left(p+\alpha h\right)\right)\Big|_{\alpha=0}= \quad \left.-\sum_{i=1}^{n}\frac{h_i/N_i}{2\left(\ln\,2\right)\left(1+\frac{P_i}{N_i}+\alpha\frac{h_i}{N_i}\right)}\quad\right._{\alpha=0},\\ &= -\sum_{i=1}^{n}\frac{h_i}{2N_i\left(\ln\,2\right)\left(1+\frac{P_i}{N_i}\right)}=-\sum_{i=1}^{n}\frac{h_i}{2\left(\ln\,2\right)\left(N_i+P_i\right)}=\left\langle\nabla f\left(p\right),h\right\rangle, \end{aligned}$$

where the gradient has entries $\left(\nabla f\left(p\right)\right)_i=-\left(2\left(\ln\,2\right)\left(N_i+P_i\right)\right)^{-1}$.

For the constraints, it is straightforward to see that $\nabla g(P)=\mathbf{1}$ and $\nabla h_i\left(P\right)=-e_i$, $i=1,...,n$.

We begin by assuming that the solution P^* is a regular point. Then the KKT conditions give that for some λ and nonnegative $\mu_1, ..., \mu_m$ we must have
Equation:

$$\begin{aligned} \sum_{i=1}^{n} \mu_i P_i^* &= 0, \\ -\frac{1}{2\left(\ln 2\right)\left(N_i + P_i^*\right)} + \lambda - \mu_i &= 0,\ i = 1, ..., n. \end{aligned}$$

The second set of constraints can be written as
Equation:

$$\begin{aligned} \frac{1}{2\left(\ln 2\right)\left(N_i + P_i^*\right)} &= \lambda - \mu_i, \\ N_i + P_i^* &= \frac{1}{2\left(\ln 2\right)\left(\lambda - \mu_i\right)}. \end{aligned}$$

Consider each inequality constraint h_i.

- If h_i is inactive, then $P_i^* > 0$ and $\mu_i = 0$. Then,
  Equation:

$$\begin{aligned} N_i + P_i^* &= \frac{1}{2\lambda(\ln 2)}, \\ P_i^* &= \frac{1}{2\lambda(\ln 2)} - N_i > 0. \end{aligned}$$

- If h_i is active, then $P_i^* = 0$ and so
  Equation:

$$\begin{aligned} N_i &= \frac{1}{2\left(\ln 2\right)\left(\lambda-\mu_i\right)}, \\ \mu_i &= \lambda - \frac{1}{2N_i\left(\ln 2\right)} \geq 0, \\ \frac{1}{2N_i\left(\ln 2\right)} &\leq \lambda, \\ \frac{1}{2\lambda(\ln 2)} &\leq N_i. \end{aligned}$$

To simplify, write $r = \frac{1}{2\lambda(\ln 2)}$; then, we have two possibilities for each channel i from above:

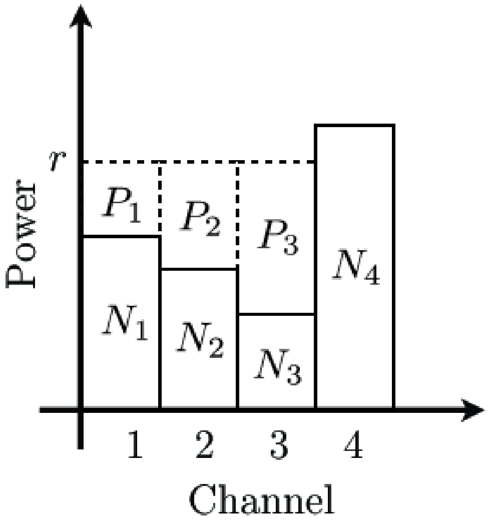
- If $r - N_i > 0$ (i.e., if $N_i < r$), then $P_i^* = r - N_i$.
- If $r - N_i \leq 0$ (i.e., if $r \leq N_i$) then $P_i^* = 0$.

Thus the power is allocated among the channels using the formula P_i^* =max $(0, r - N_i)$, and the value of r is chosen so that the total power constraints is met:

Equation:

$$\sum_{i=1}^{n} \max\left(0, r - N_i\right) = P_T.$$

This is the famous *water-filling solution* to the multiple channel capacity problem, illustrated in [link].



Waterfilling solution to the multiple channel power allocation problem, which is solved using Karush-Kuhn-Tucker conditions.

Fenchel Duality
Introduces the concepts of an epigraph and a conjugate function necessary to set up Fenchel dual problems.

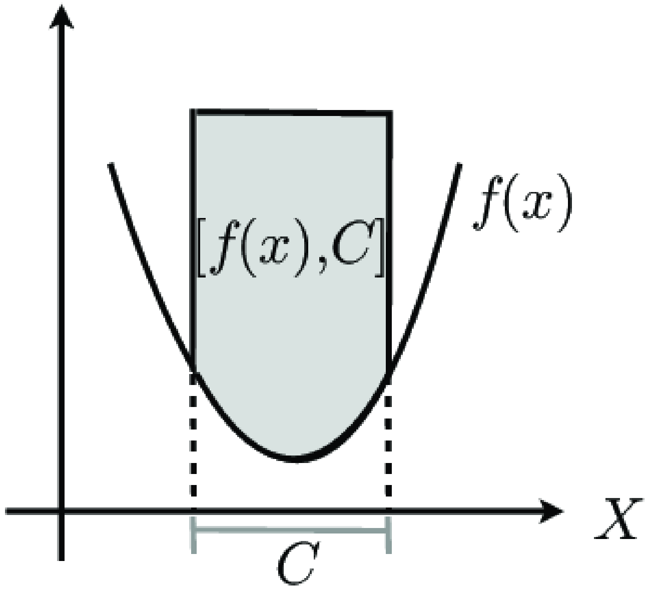
Convexity and Conjugate Functions

We begin by reviewing two notions of convexity: for sets and for functions.

Definition 1 A subset $C \subseteq X$ is called *convex* if $z = \alpha x + (1 - \alpha) y \in C$ for every $x, y \in C$ and $\alpha \in [0, 1]$; z is called a *convex combination* of x and y.

Definition 2 Let C be a convex set. A functional $f : C \to \mathbb{R}$ is *convex onC* if $f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y)$ for all $x, y \in C$ and $\alpha \in [0, 1]$. If strict inequality holds the functional is set to be *strictly convex*. A functional g is called *concave (strictly concave)* if $-g$ is convex (strictly convex).

We will denote the region above the function f defined over a convex set C as $[f, C]$, sometimes called an *epigraph*, as illustrated in [link].



Example of an epigraph.

Definition 3 Let f be a convex functional on a convex set C. The *conjugate set* C^* is defined as

Equation:

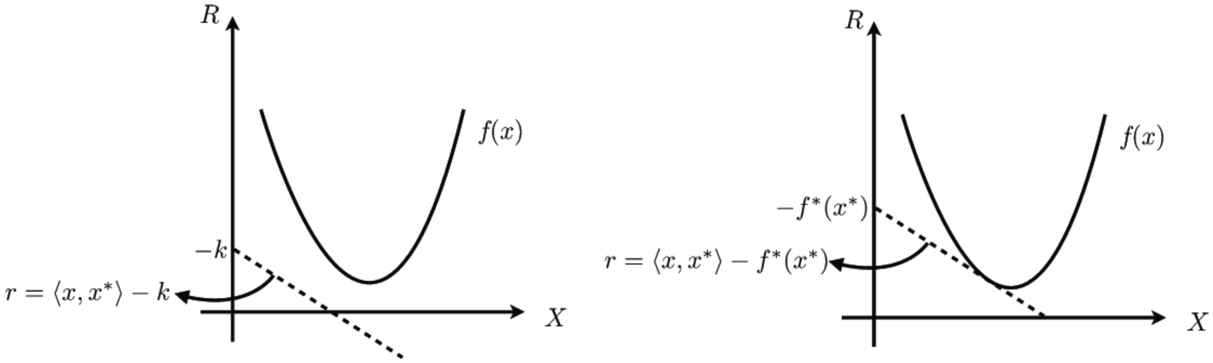
$$C^* = \left\{x^* \in X : \sup_{x \in C} \left[\left\langle x, x^* \right\rangle - f(x)\right] < \infty\right\},$$

and the *conjugate functional* $f^* : C^* \rightarrow \mathbb{R}$ is defined as

Equation:

$$f^*\left(x^*\right) = \sup_{x \in C} \left[\left\langle x, x^* \right\rangle - f(x)\right].$$

There is a geometric intuition behind the definition of the conjugate functionals. Consider the illustration below where the horizontal axis represents the space X and the vertical axis represents the scalar field. A hyperplane in this space contains all points $(x, r) \in X \times R$ for which $r = \left\langle x, x^* \right\rangle - k$ for some value of $k \in R$; the vector x^* determines the orientation of the hyperplane and the value k determines the shift from the origin (i.e., the intersect in the axis R). The value of the functional $f^*\left(x^*\right)$ corresponds to the supremum value of k for which the hyperplane intersects $[f, C]$, and is finite only for $x^* \in C^*$; this is illustrated in [link].


R
$f(x)$
$-k$
$r = \langle x, x^* \rangle - k$
X
R
$f(x)$
$-f^*(x^*)$
$r = \langle x, x^* \rangle - f^*(x^*)$
X

The conjugate function of a convex epigraph.

Note that C^* is convex and f^* is convex. This definition is easily extended to concave functionals.

Definition 4 Let g be a concave functional on a convex set D. The *conjugate set* D^* is defined as
Equation:

$$D^* = \left\{x^* \in X : \inf_{x\in D} \left[\left\langle x, x^*\right\rangle - g\left(x\right)\right] > -\infty\right\},$$

and the *conjugate functional*$g^* : D^* \to \mathbb{R}$ is defined as
Equation:

$$g^*\left(x^*\right) = \inf_{x\in D} \left[\left\langle x, x^*\right\rangle - f\left(x\right)\right].$$

Note that D^* is convex and g^* is concave.

Fenchel Duality

The following theorem will allow us to convert an optimization problem with a convex objective function into a dual problem with a concave objective function.

Theorem 1 (Fenchel) Assume that f and g are convex and concave functions, respectively, on convex sets C and D in a normed space X. Assume that $C \cap D$ contains points in the relative interior of C and D and that either $[f, C]$ or $[g, D]$ has a nonempty interior. Suppose further that
Equation:

$$\mu = \inf_{x\in C\cap D} \left\{f\left(x\right) - g\left(x\right)\right\}$$

is finite. Then,
Equation:

$$\mu = \inf_{x\in C\cap D} \left\{f\left(x\right) - g\left(x\right)\right\} = \max_{x^*\in C^*\cap D^*} \left\{g^*\left(x^*\right) - f^*\left(x^*\right)\right\},$$

where the maximum is achieved by some $x_0^* \in C^* \cap D^*$.

In this theorem, $g(x)$ is usually set to zero. From a geometrical point of view, the theorem states that there are two ways to interpret the minimum distance between the two epigraphs $[f, C]$ and $[g, D]$ shown below: one in terms of the original functions f, g and one in terms of the duals f^*, g^*: we look for the two tangent hyperplanes for f and g that are maximally separated from one another, as illustrated in [link].

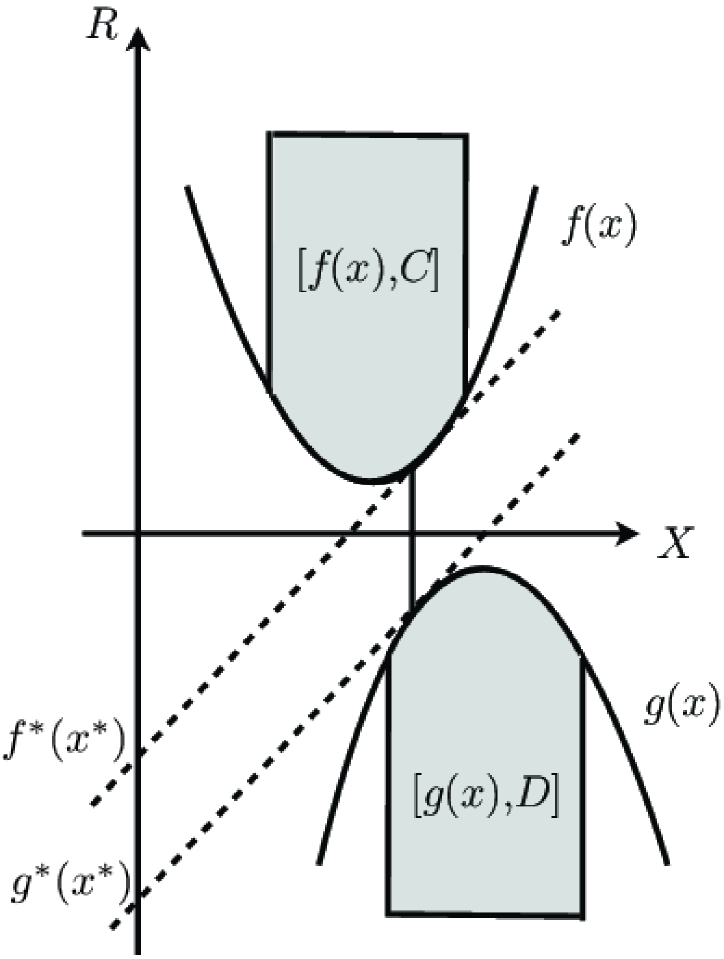


Illustration of the Fenchel dual problem on a conjugate function.

If the infimum on the left is achieved by x_0, then
Equation:

$$\max_{x \in C} \left[\left\langle x, x_0^* \right\rangle - f(x) \right] = \left\langle x_0, x_0^* \right\rangle - f(x_0),$$
$$\max_{x \in D} \left[\left\langle x, x_0^* \right\rangle - g(x) \right] = \left\langle x_0, x_0^* \right\rangle - g(x_0).$$

Example 1 (Allocation) Assume that we have a capital x_0 available for investment with n different funds. There is a predicted gain $g_i(x_i)$ to having stock worth x_i at fund i, where the functions g_i are concave. We aim to find the optimal allocation of the capital $x = (x_1, ..., x_n)$ that maximizes the total gain $g(x) = \sum_{i=1}^{n} g_i(x_i)$.

To appeal to duality, we have the concave function $g(x)$ and must define a convex function, e.g., $f(x) = 0$. The constraint set can be written as the intersection $C \cap D$ with
Equation:

$$\begin{aligned} C &= \left\{ x : \sum_{i=1}^{n} x_i = x_0 \right\} = \{ x : \langle \mathbf{1}, x \rangle = x_0 \}, \\ D &= \{ x : x_i \geq 0, i = 1, ..., n \}. \end{aligned}$$

Therefore, we can write our optimization problem as
Equation:

$$\min_{x \in C \cap D} \{ f(x) - g(x) \}.$$

We consider the conjugate sets. First, we have
Equation:

$$\begin{aligned} C^* &= \left\{ x^* \in X : \sup_{x \in C} \left[\left\langle x, x^* \right\rangle - f(x) \right] < \infty \right\}, \\ &= \left\{ x^* \in X : \sup_{x \in C} \left[\left\langle x, x^* \right\rangle \right] < \infty \right\}. \end{aligned}$$

We want to define C^{*} more explicitly. Let $x^{*} \in C^{*}$ be written as $x^{*} = \lambda\mathbf{1} + w$, where $w \perp \mathbf{1}$. Then,

Equation:

$$\left\langle x^{*}, w \right\rangle = \lambda \left\langle \mathbf{1}, w \right\rangle + \left\langle w, w \right\rangle = \| \ w \ \|^{2},$$

which can be arbitrarily large. Now let $x = \frac{x_0}{n}\mathbf{1} + \lambda w$; it is easy to check that $x \in C$ for all $\lambda \in \mathbb{R}$. If $w \neq 0$ then $\left\langle x, x^{*} \right\rangle = \lambda x_0 + \lambda \| \ w \ \|^{2}$, which again can be arbitrarily large. Since $x^{*} \in C^{*}$ must hold that $\sup_{x \in C} \left\langle x, x^{*} \right\rangle < \infty$, we must have that $w = 0$ and so $x^{*} \in \text{span}\left(\{\mathbf{1}\}\right)$. Since $x^{*} \in C^{*}$ was arbitrary, then $C^{*} \subseteq \text{span}\left(\{\mathbf{1}\}\right)$.

It is also easy to see that $\text{span}\left(\{\mathbf{1}\}\right) \subseteq C^{*}$, therefore implying that $C^{*} = \text{span}\left(\{\mathbf{1}\}\right)$.

For D, we have a conjugate set

Equation:

$$\begin{aligned} D^{*} \ &= \left\{ y : \inf_{x \in D} \left[\left\langle x, y \right\rangle - g\left(x\right) \right] > -\infty \right\}, \\ &= \left\{ y : \inf_{x \in D} \left[\left\langle x, y \right\rangle - g\left(x\right) \right] > -\infty \right\}, \end{aligned}$$

since $g\left(x\right) \leq x_0$. Now $D \subseteq D^{*}$ since all vectors in D have nonnegative entries. Fix $y \in D^{*}$; if $y \notin D$ then there is some negative entry $y_i < 0$ among $i = 1, ..., n$. For such i let $x_\lambda = \lambda e_i \in D$ for some $\lambda > 0$; then we get $\left\langle x_\lambda, y \right\rangle = \lambda y_i$ which can be arbitrarily close to ∞ (i.e., as $\lambda \rightarrow \infty$ we have $\left\langle x_\lambda, y \right\rangle \rightarrow -\infty$. Thus $y \notin D^{*}$, a contradiction. Therefore, if $y \in D^{*}$ then $y \in D$ and so $D^{*} \subseteq D$. We have therefore shown that $D^{*} = D$.

The conjugate functionals can be written as

Equation:

$$f^{*}\left(x^{*}\right) = \sup_{x \in C} \left\langle x, x^{*} \right\rangle = \lambda x_0,$$

since each $x^* \in C^*$ can be written as $x^* = \lambda \mathbf{1}$. Therefore, f^* can be written as a function of a single variable. Similarly,

Equation:

$$g^*\left(x^*\right) = \inf_{x \in D} \left(\left\langle x, x^*\right\rangle - g\left(x\right)\right) = \inf_{x \in D} \left(\sum_{i=1}^{n} x_i x_i^* - g_i\left(x_i\right)\right) = \sum_{i=1}^{n} g_i^*\left(x_i^*\right),$$

where we write

Equation:

$$g_i^*\left(x_i^*\right) = \inf_{x_i > 0} \left(x_i x_i^* - g_i\left(x_i\right)\right).$$

For $x^* \in C^* \cap D^*$ we can write $x_i = \lambda > 0$ for all $i = 1, \ldots, n$ and so

Equation:

$$g_i^*\left(x_i^*\right) = g_i^*\left(\lambda\right) = \inf_{x_i > 0} \left(\lambda x_i - g_i\left(x_i\right)\right).$$

Therefore, the original problem can be reformulated as the following single-variable problem:

Equation:

$$\lambda^* = \min_{\lambda > 0} \left[\lambda x_0 - \sum_{i=1}^{n} g_i^*\left(\lambda\right)\right].$$

This is due to $x^* = \lambda \mathbf{1} \in C^* \cap D^*$ if and only if $\lambda = 0$. Once λ^* is found, we can find each x_i as the minimizer in $g_i^*\left(\lambda^*\right)$, cf. [link].